Two-Stage Document Length Normalization for Information Retrieval

SEUNG-HOON NA, Busan University of Foreign Studies, South Korea

The standard approach for term frequency normalization is based only on the document length. However, it does not distinguish the verbosity from the scope, these being the two main factors determining the document length. Because the verbosity and scope have largely different effects on the increase in term frequency, the standard approach can easily suffer from insufficient or excessive penalization depending on the specific type of long document. To overcome these problems, this paper proposes two-stage normalization by performing verbosity and scope normalization separately, and by employing different penalization functions. In verbosity normalization, each document is pre-normalized by dividing the term frequency by the verbosity of the document. In scope normalization, an existing retrieval model is applied in a straightforward manner to the pre-normalized document, finally leading us to formulate our proposed verbosity normalized (VN) retrieval model. Experimental results carried out on standard TREC collections demonstrate that the VN model leads to marginal but statistically significant improvements over standard retrieval models.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*

General Terms: Algorithms, Experimentation, Performance, Theory

Additional Key Words and Phrases: verbosity normalization, scope normalization, document length normalization, retrieval heuristics, term frequency

ACM Reference Format:
Seung-Hoon Na, 2014. Two-stage document length normalization for information retrieval. *ACM Trans. Inf. Syst.* 0, 0, Article 00 (2014), 39 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

In information retrieval (IR), term frequency is a fundamental and important component of a ranking model. Intuitively, the larger the term frequency of a query word in a document, the more likely the document is to be about the query topic, and thus, the document should have a higher relevance score. In practice, however, documents are of various lengths, and the simple approach of preferring documents with higher term frequency could easily result in an excessive preference for long documents. To use the term frequency in a fairer approach, *normalization* of the term frequency has been extensively investigated by researchers.

With regard to the normalization problem, Robertson and Walker introduced the verbosity and the scope hypotheses, which state that document length is mainly determined by two factors – *verbosity* and *scope* – as follows [Robertson and Walker 1994; Robertson and Zaragoza 2009]:

Author e-mail: nash@bufs.ac.kr
This work was partly supported by the IT R&D program of MSIP/KEIT. [10041807, Development of Original Software Technology for Automatic Speech Translation with Performance 90% for Tour/International Event focused on Multilingual Expansibility and based on Knowledge Learning] and by the research grant of the Busan University of Foreign Studies in 2014.

DOI: http://dx.doi.org/10.1145/0000000.0000000

1) **Verbosity hypothesis**: "*Some authors are simply more verbose, using more words to say the same thing* [Robertson and Zaragoza 2009]."

2) **Scope hypothesis**: "*Some authors have more to say: they may write a single document containing or covering more ground* [Robertson and Zaragoza 2009]."

In this paper, we focus on the difference between the effect of the verbosity and the scope on the term frequency of a single word. Verbosity, as the name implies, is related to the burstiness of term frequency, which helps an already mentioned word in a document get a higher frequency. Even if a word has a low term frequency in normal verbosity, its term frequency could increase significantly when the document has high verbosity. On the other hand, scope mostly involves the creation of a new word, rather than boosting the term frequency. Broadening the scope of a document would help unseen words in a normal document get non-zero frequencies. However, these non-zero frequencies might not be high. Therefore, verbosity leads to a significant increase in term frequency, whereas scope leads to a rather limited increase in term frequency. In other words, the scope of a document only helps the occurrence of a new word, and the term frequency of the word is mostly governed by the verbosity of the document.

Despite this difference between verbosity and scope, standard normalization is a length-driven approach, i.e., it is based only on the document length, without distinguishing between verbosity and scope. As a result, it may suffer from *insufficient* penalization of a *verbose* document whose length is increased mainly by high verbosity, and *excessive* penalization of a *broad* document whose length is mainly derived from the broad scope.

In the light of this addressed difference, this paper argues that verbosity and scope should be normalized separately by employing different penalization functions. To achieve this, we propose a two-stage normalization approach. We first perform *verbosity normalization* for each document by linearly dividing the term frequency by the verbosity, thus obtaining a *verbosity-normalized document representation*. We then perform *scope normalization*, in which an existing retrieval model is applied to this verbosity-normalized document representation. The final model obtained is called a *verbosity-normalized* (VN) *retrieval model*.

Furthermore, we examine whether the proposed VN retrieval model resulting from two-stage normalization performs the desired separate penalizations. Toward this end, we first select three popular retrieval models – the Okapi model [Robertson et al. 1995], the Dirichlet-prior (DP) smoothed language model [Zhai and Lafferty 2001], and the Markov random field (MRF) model [Metzler and Croft 2005] – and then perform *comparative axiomatic analysis* of the original and the VN retrieval models, under the setting of the axiomatic framework introduced in [Fang et al. 2004; Fang et al. 2011]. The analysis results confirm that the VN model indeed performs the desired separate normalizations, i.e., a *strict* penalization of verbosity-increased documents and a *relaxed* penalization of scope-broadened documents.

The results of experiments carried out on standard TREC test collections show that the VN retrieval models are significantly better than the original models. The experimental results support our motivating argument that the verbosity and scope should be handled separately using different penalization functions.

The remainder of this paper is organized as follows. Section 2 describes previous studies. Section 3 describes the proposed two-stage normalization approach and presents the VN retrieval models for DP, Okapi, and MRF. Section 4 presents the main results of the analysis of retrieval models under standard length normalization constraints. Sections 5, 6, 7 present the experimental setting and results. Sections 8 concludes.

2. PREVIOUS WORK

Singhal et al. [1996] recognized that simply dividing the term frequency by the document length leads to the over-penalization problem in long documents. To overcome this problem, they proposed *pivoted normalization*, in which a pivoted length is used to normalize the term frequency by adding a constant pivot factor (i.e., average document length) to the original document length. Pivoted normalization had originally been introduced in Okapi's model [Robertson et al. 1995], before it was formalized and named by Singhal et al. [1996]. Because pivoted normalization yields successful results, it has been explicitly adopted by other retrieval models, such as the INQUERY system [Callan et al. 1992; Allan et al. 2000]. A similar relaxed type of normalization has been commonly used in more recent retrieval models – normalization 2 in the divergence from randomness (DFR) retrieval framework [Amati and Van Rijsbergen 2002] and the smoothed document length in DP [Zhai and Lafferty 2001].

Fang et al. [2004] formally and mathematically defined IR heuristics, drawn from ranking characteristics most commonly used by existing retrieval models, thereby proposing a novel direction for an axiomatic approach to IR. The retrieval heuristics defined in the axiomatic approach have been used to define a new retrieval model inductively [Fang and Zhai 2005; Clinchant and Gaussier 2010] and to restrict the search space for automatically learning a retrieval function [Cummins and O'Riordan 2006]. In addition to original constraints, some studies have explored new constraints including: semantic term matching constraints [Fang and Zhai 2006], the proximity-based matching constraint [Tao and Zhai 2007], the burstness-based normalization constraint [Clinchant and Gaussier 2010], the document frequency constraint for pseudo-relevance feedback [Clinchant and Gaussier 2011], the feedback term weight constraints for pseudo-relevance feedback [Clinchant and Gaussier 2013], and the translation probability constraints for translation language models [Karimzadehgan and Zhai 2012]. With regard to the length normalization problem, Fang et al. [2004] defined three length normalization constraints (referred to as LNC1, LNC2, and TF-LNC), demonstrating analytically that popular retrieval models satisfy all these normalization constraints at least for content-bearing words.

Our argument that different normalization functions should be used for verbosity and scope was also proposed by [Robertson and Walker 1994; Robertson and Zaragoza 2009], in a more restricted manner, as follows: "*The verbose hypothesis suggests that we should simply normalize term frequencies by dividing by document length, while the scope hypothesis, on the other hand, suggests the opposite* [Robertson and Zaragoza 2009]." That is, they suggest that a retrieval function does not necessarily penalize a long document when it has a broad scope. A similar argument was also made by [Na et al. 2008a]. Our suggestion, however, is that we still need the penalization for scope, but in a much more relaxed manner. In this sense, our argument can be regarded as a generalization of the previous arguments.

To the best of our knowledge, one of the first approaches for two-stage normalization is *pivoted unique normalization*, suggested by [Singhal et al. 1996]. In their approach, the term frequency is first normalized on the basis of a nonlinear function by using the average term frequency (which corresponds to verbosity normalization), and the normalized term frequency is then further divided by a pivoted unique length (which corresponds to scope normalization). However, it remains unclear how their approach can be generalized to other retrieval models.

Going beyond the aforementioned existing works, we propose a generalized two-stage normalization approach, arguing more clearly that the term frequency should be penalized differently, depending on whether a document is long because of the

verbosity or the scope. Our approach is not limited to a specific retrieval model or a specific measure of the verbosity or scope. We also analytically present the retrieval heuristics realized by two-stage normalization, by performing a comparative axiomatic analysis under the setting of standard normalization constraints suggested by Fang et al. [2004],

It is noteworthy that the Okapi and the DFR retrieval framework [Amati and Van Rijsbergen 2002] can be considered as another type of two-stage normalization. According to the derivation by [He and Ounis 2003], the first step normalizes the term frequency by a relaxed document length using $tfn = c(w,d)/\left(k_1\left((1-b)+b\cdot|d|/avgl\right)\right)$ in Okapi and $tfn = c(w,d)log\left(1+c\cdot avgl/|d|\right)$ in DFR, and the second step further normalizes tfn by $tfn/(tfn+1)$. The first step uses the document length, thereby performing a mixed normalization of verbosity and scope, and the second step roughly performs verbosity normalization by preventing a document with high tfn from getting a very large score. However, this is not the case in our approach, which further distinguishes between the verbosity and the scope.

Interestingly, passage retrieval can also be viewed as two-stage normalization [Salton and Buckley 1991; Salton et al. 1993; Callan 1994; Allan 1995; Mittendorf and Schäuble 1994; Kaszkiel and Zobel 1997; Kaszkiel et al. 1999; Liu and Croft 2002; Bendersky and Kurland 2008; Na et al. 2008b; Bendersky and Kurland 2010; Lv and Zhai 2009b; Lv and Zhai 2010; Krikon et al. 2010; Krikon and Kurland 2011]. Because scopes are more similar in passages themselves than in documents, using passages itself can be considered as a type of scope normalization. Thereafter, applying an existing retrieval method to score each passage corresponds to verbosity normalization.

Recently, Lv and Zhai [2011c] and Lv and Zhai [2011b] observed that when documents are extremely long, the score gap calculated as the difference between scores when a query term is present and when it is absent in a document, could be infinitely close to zero or negative. As a result, extremely long documents tend to be overly penalized. To ensure a desirable score gap between documents that match and do not match a query term, Lv and Zhai [2011b] proposed *lower-bounding* term frequency normalization, which can be described as follows: (1) A *pseudo score gap* between documents that match and do not match a query term is newly introduced as a document-independent factor. (2) For each query term, the pseudo score gap is added to the original document score only when the document matches the query term, whereas the original document score is left unchanged for a document that does not match the query term[1]. Importantly, Lv and Zhai [2011b] closely examined the underlying principles of their proposed normalization, after which they proposed the constraints **LB1** and **LB2** as extensions of the existing formal heuristics used in [Fang et al. 2011]. According to their axiomatic analysis, all modified retrieval functions proposed in [Lv and Zhai 2011b] unconditionally or more easily satisfy the lower bounds (LBs) without violating the original constraints of [Fang et al. 2011], whereas existing functions do not satisfy the LBs. Experiment results showed that all modified retrieval functions showed statistically significant improvements, especially for verbose queries. In contrast to our work, the lower-bounding normalization proposed in [Lv and Zhai 2011b] uses only the document length. However, in our case, we distinguish the verbosity of the document from the scope. In addition, the new constraints used in [Lv and Zhai 2011b] are complementary to the existing length normalization

[1]The same scoring function can be equivalently implemented by redefining a within-document scoring function for both cases (i.e., either a document matches a query or it does not), as formulated in [Lv and Zhai 2011b].

constraints (LNCs), whereas our work emphasizes the need to pursue a new generation of LNCs.

2.1. Novel Contributions beyond Our Prior Work

In [Na et al. 2008a], the initial form of the two-stage normalization approach was presented to modify language modeling approaches by introducing the *pseudo document model*. However, [Na et al. 2008a] were not aware of the importance of the pseudo document model as a generalized solution for handling the addressed problem. In addition, [Na et al. 2008a] suggested a rather harsh retrieval constraint called TNC, which is too strong to be satisfied by even their own proposed method. Given the previous presentation of [Na et al. 2008a], it thus remained unclear how the presented normalization yields some of the reported improved performances, and how it can be generalized to other retrieval models. Building on our prior work, novel contributions of this paper are listed in the following:

— *Generalized* two-stage normalization (Section 3), which was not explicitly argued and not fully formalized in [Na et al. 2008a]. With the explicit formulation, we now correctly understand VN-DP as a specific instance of two stage normalization.
— Extensions to other models – Okapi and MRF (Section 3), as a result of the proposed generalized normalization
— Analytically capturing retrieval heuristics of two-stage normalization by performing *comparative axiomatic analysis* (Section 4 & Appendices C, D, and E) under the standard constraint setting of [Fang et al. 2004; Fang et al. 2011]
— *LengthPower* as a novel scope measure (Section 3). Using LengthPower, we have an unified view of language modeling approaches by considering both JM and DP as special cases of VN-DP.
— *Comparison with lower bounding term frequency normalization* (Section 7)

3. TWO-STAGE NORMALIZATION

In this section, we describe our proposed two-stage normalization in detail, and apply it to the DP, Okapi, and MRF approaches, as case studies.

3.1. Verbosity Normalization

The following are notations commonly used in this paper.
• V: $w_1, w_2, \cdots, w_{|V|}$, Set of all words
• N: Number of documents in a given collection
• C: A given collection, consisting of $d_1, \cdots, d_N$. Often, we also use C to refer to the concatenated representations of all documents in C.
• $df(w)$: Document frequency of w
• d (or q): A given document (or a query)
• $c(w, d)$ (or $c(w, q)$): Term frequency of word w in document d (or query q)
• $c(w, C)$: Term frequency of word w in collection C defined by $\sum_{d \in C} c(w, d)$
• $idf(w)$: Term discrimination value of w such as IDF
• $|d|$: Length of document d, defined by $\sum_{w \in V} c(w, d)$
• $|C|$: Length of collection C, defined by $\sum_{w \in V} c(w, C)$ (for brevity of notation, C is either the set of documents or the concatenated representation of documents, depending on context)
• $s(d)$: Scope of document d ($s(d) \leq |d|$)
• $v(d)$: Verbosity of document d
• $avgl, avgv, avgs$: Average length, verbosity, and scope, respectively, of documents in the collection.

Motivated by the verbosity and the scope hypotheses, we first assume that the document length is decomposed into the verbosity and the scope, thereby providing the following simplified formula:

$$|d| = v(d)s(d) \tag{1}$$

As a result, we can formulate $v(d)$ in terms of $s(d)$ and $|d|$ as follows:

$$v(d) = \frac{|d|}{s(d)} \tag{2}$$

The derivation of Eq. (2) is presented in Appendix A.

In verbosity normalization, the original term frequency is normalized by dividing it by the verbosity of the document. To formally describe verbosity normalization, let ϕ be a *verbosity normalization operator*; $\phi(d)$, the *verbosity-normalized document representation* of d[2], which is the document transformed by applying the operator ϕ to all words in a document d; and $c(w, \phi(d))$, the *verbosity-normalized term frequency* of word w. Then, verbosity normalization refers to the process of obtaining $c(w, \phi(d))$ for word w, using the following formula:

$$c(w, \phi(d)) = k\frac{c(w, d)}{v(d)} \tag{3}$$

where k is a verbosity scaling parameter. By substituting Eq. (2) into Eq. (3), $c(w, \phi(d))$ becomes

$$c(w, \phi(d)) = k\frac{c(w, d) \cdot s(d)}{|d|}$$

The resulting normalized term frequency is not only inversely proportional to the document length but is also proportional to the scope of the document.

3.2. Scope Normalization

For scope normalization, we need to consider a more relaxed function than that for verbosity normalization. We first note that the scope of an original document is the verbosity-normalized length of the document, as follows:

$$|\phi(d)| = \sum_{w \in V} c(w, \phi(d)) = k\frac{\sum_{w \in V} c(w, d) \cdot s(d)}{|d|} = k \cdot s(d)$$

Furthermore, existing retrieval models perform a type of relaxed normalization by using their pivoted length or smoothed length. Thus, instead of developing a new function, we perform scope normalization by straightforwardly applying an existing retrieval model to the verbosity-normalized document representation $\phi(d)$. Formally, let $f(d, q)$ be the original retrieval function that gives a score to d, for query q. Applying two-stage normalization to $f(d, q)$ gives $f(\phi(d), q)$, which is obtained by replacing $c(w, d)$ used in all terms in $f(d, q)$ with $c(w, \phi(d))$ for all documents in the collection. We call $f(\phi(d), q)$ a **VN** (verbosity-normalized) **retrieval model** or a **VN scoring function**.

[2]In our notation, the verbosity normalization operator is applied not to document itself but instead to the document representation. In this paper, the document representation is assumed to be a vector of its term frequencies (and either bigram or proximal term frequencies). For general purposes, the verbosity normalized operator needs to be extended such that it can be applied to advanced document representation such as a sequence of words, so that it can be useful for the proximity-based or location-based search.

3.3. Examples of Verbosity-Normalized Retrieval Models

In this section, we present the application of two-stage normalization to the DP, Okapi, and MRF approaches.

3.3.1. Dirichlet-prior (DP). DP performs Bayesian smoothing on a multinomial language model [Zhai and Lafferty 2001], for which the conjugate prior is the Dirichlet distribution with the following parameters:

$$\big(\mu p(w_1|C), \mu p(w_2|C), \cdots, \mu p(w_{|V|}|C)\big) \tag{4}$$

The Bayesian priors using the parameters of Eq. (4) give the following smoothed model of document d:

$$P(w|\phi(d)) = \frac{c(w,d) + \mu p(w|C)}{|d| + \mu}$$

and the following scoring function for a given query q [Zhai and Lafferty 2001]:

$$\sum_{w \in q \cap d} ln\left(1 + \frac{c(w,d)}{\mu \cdot p(w|C)}\right) + |q| \cdot ln\left(\frac{\mu}{|d| + \mu}\right)$$

The VN model $f(\phi(d), q)$ is assumed to employ the following *document-specific* conjugate prior:

$$\big(\mu v(d) p(w_1|C), \mu v(d) p(w_2|C), \cdots, \mu v(d) p(w_{|V|}|C)\big) \tag{5}$$

In other words, the more verbose d is, the larger is the prior probability used. A detailed justification for Eq. (5) is presented in Appendix B. These modified Bayesian priors using the parameters of Eq. (5) give the following smoothed model:

$$P(w|d) = \frac{c(w,d) + \mu v(d) p(w|C)}{|d| + \mu v(d)} \tag{6}$$

We simply use $k = 1$ in Eq. (3), because the scaling parameter k of $c(w, \phi(d))$ is absorbed into the smoothing parameter μ. Then, Eq. (6) becomes

$$P(w|\phi(d)) \;=\; \frac{c(w,\phi(d)) + \mu \cdot p(w|C)}{|\phi(d)| + \mu} \tag{7}$$

Eq. (7) is the same as the equation obtained by replacing $c(w,d)$ with $c(w,\phi(d))$.

Using Eq. (6), the resulting retrieval function is given as

$$\sum_{w \in q \cap d} c(w,q) ln\left(1 + \frac{c(w,d)}{\mu \cdot p(w|C)} \frac{s(d)}{|d|}\right) + |q| \cdot ln\left(\frac{\mu}{s(d) + \mu}\right)$$

which is called **VN-DP**[3].

3.3.2. Okapi. Okapi's BM25 retrieval formula, as presented by [Robertson et al. 1995], is

$$\sum_{w \in q \cap d} \left\{ \frac{(k_3+1)c(w,q)}{k_3 + c(w,q)} ln\left(\frac{N - df(w) + 0.5}{df(w) + 0.5}\right) tf_{BM25}(w,d) \right\}$$

where the term frequency component $tf_{BM25}(w,d)$ is

$$tf_{BM25}(w,d) = \frac{(k_1+1)c(w,d)}{k_1\left((1-b) + b\frac{|d|}{avgl}\right) + c(w,d)}$$

[3]The formula of VN-DP is equivalent to the modified Dirichlet-prior smoothing suggested by [Na et al. 2008a].

Table I. Feature functions used in the MRF model. $c_{\#1}(q_iq_{i+1}, d)$ indicates the number of times that the *exact phrase* q_iq_{i+1} occurs in document d, and $c_{\#un8}(q_iq_{i+1}, d)$ indicates the number of times that both terms q_i and q_{i+1} appear *ordered* or *unordered* within a window with a span of 8.

Feature	Value
$f_T(d, q_i)$	$\ln \left[\frac{c(q_i,d)+\mu_T \frac{c(q_i,C)}{\|C\|}}{\|d\|+\mu_T} \right]$
$f_O(d, q_iq_{i+1})$	$\ln \left[\frac{c_{\#1}(q_iq_{i+1},d)+\mu_O \frac{c_{\#1}(q_iq_{i+1},C)}{\|C\|}}{\|d\|+\mu_O} \right]$
$f_U(d, q_iq_{i+1})$	$\ln \left[\frac{c_{\#un8}(q_iq_{i+1},d)+\mu_U \frac{c_{\#un8}(q_iq_{i+1},C)}{\|C\|}}{\|d\|+\mu_U} \right]$

Here, k_1, k_3, and b are constants. In the VN model, the IDF part is not changed; however, $tf_{BM25}(w, d)$ is modified to $tf_{BM25}(w, \phi(d))$ obtained by replacing $c(w, d)$ with $c(w, \phi(d))$, as follows:

$$tf_{BM25}(w, \phi(d)) = \frac{(k_1 + 1)c(w, \phi(d))}{k_1 \left((1 - b) + b\frac{|\phi(d)|}{avgs} \right) + c(w, \phi(d))}$$

As in the case of DP, we assume the scale parameter k to be 1, because it is absorbed into k_1, resulting in the following final form:

$$tf_{BM25}(w, \phi(d)) \; = \; \frac{(k_1 + 1)c(w, d)}{k_1|d| \left((1 - b)\frac{1}{s(d)} + b\frac{1}{avgs} \right) + c(w, d)}$$

The modified Okapi function by using $tf_{BM25}(w, \phi(d))$ for $tf_{BM25}(w, d)$ is called **VN-Okapi**.

3.3.3. Markov Random Field (MRF). MRFs are undirected graphical models that are used to define joint distributions over a set of random variables. The use of MRFs for IR was suggested by [Metzler and Croft 2005; Metzler and Bruce Croft 2007], going beyond the simplistic bag of words assumption, by explicitly modeling the term dependency among query words. Thus far, three different variants of the MRF model have been suggested according to the type of dependency assumed among query words – full independence, sequence dependence, and full dependence. This paper focuses on *sequence dependence*, which has been widely used in many recent works [Metzler and Croft 2007; Lease 2009; Bendersky et al. 2010; Wang et al. 2010; Lang et al. 2010; Bendersky et al. 2011], because of its good balance between effectiveness and efficiency.

To formally present the ranking function of the sequential dependence, suppose that q is a sequence of m terms $q_1 \cdots q_m$. According to the original framework, the relevance score of a document d is given by [Metzler and Croft 2005]

$$f(d, q) = \lambda_T \sum_{q_i \in q} f_T(d, q_i) + \lambda_O \sum_{q_iq_{i+1} \in q} f_O(d, q_iq_{i+1}) + \lambda_U \sum_{q_iq_{i+1} \in q} f_U(d, q_iq_{i+1}) \qquad (8)$$

where we have the constraint $\lambda_T + \lambda_O + \lambda_U = 1$, and $f_T(d, q_i)$, $f_O(d, q_iq_{i+1})$ and $f_U(d, q_iq_{i+1})$ are called the *feature functions* of the term, *ordered phrase*, and *unordered phrases*, respectively. Table I presents the definition of each feature function [Metzler and Croft 2005].

Table II. Verbosity-normalized feature functions used in the VN-MRF model.

Feature	Value
$f_T(\phi(d), q_i)$	$\ln\left[\frac{\frac{c(q_i,d)}{v(d)}+\mu_T\frac{c(q_i,C)}{\lvert C\rvert}}{s(d)+\mu_T}\right]$
$f_O(\phi(d), q_iq_{i+1})$	$\ln\left[\frac{\frac{c_{\#1}(q_iq_{i+1},d)}{v(d)}+\mu_O\frac{c_{\#1}(q_iq_{i+1},C)}{\lvert C\rvert}}{s(d)+\mu_O}\right]$
$f_U(\phi(d), q_iq_{i+1})$	$\ln\left[\frac{\frac{c_{\#un8}(q_iq_{i+1},d)}{v(d)}+\mu_U\frac{c_{\#un8}(q_iq_{i+1},C)}{\lvert C\rvert}}{s(d)+\mu_U}\right]$

Following the original framework [Metzler and Croft 2005], we assume that μ_T, μ_O, and μ_U are the same, i.e., $\mu_T = \mu_O = \mu_U = \mu$, unless otherwise stated. We refer to the retrieval function in Eq. (8) as **MRF**.

To derive a VN retrieval model $f(\phi(d), q)$ for MRF, we replace the original term frequencies with the verbosity normalized ones. For this purpose, let $c_{\#1}(q_iq_{i+1}, \phi(d))$ and $c_{\#un8}(q_iq_{i+1}, \phi(d))$ be *VN ordered* and *unordered phrase term frequencies* for q_iq_{i+1}, respectively. Similar to the definition of VN term frequency in Eq. (3), these VN phrase term frequencies are defined as follows:

$$c_{\#1}(q_iq_{i+1}, \phi(d)) = k\frac{c_{\#1}(q_iq_{i+1}, d)}{v(d)} \tag{9}$$

$$c_{\#un8}(q_iq_{i+1}, \phi(d)) = k\frac{c_{\#un8}(q_iq_{i+1}, d)}{v(d)} \tag{10}$$

Furthermore, let $f_T(\phi(d), q_i)$, $f_O(\phi(d), q_iq_{i+1})$, and $f_U(\phi(d), q_iq_{i+1})$ be *VN feature functions* that correspond to original feature functions. Table II describes the definition of each VN feature function. As in the case of VN-DP, k is assumed to be 1 in all VN feature functions, because it is absorbed to μ_T, μ_O, or μ_U. Finally, we obtain the scoring function for the VN model $f(\phi(d), q)$ of MRF as follows:

$$f(\phi(d), q) = \lambda_T \sum_{q_i \in q} f_T(\phi(d), q_i) + \lambda_O \sum_{q_iq_{i+1} \in q} f_O(\phi(d), q_iq_{i+1}) + \lambda_U \sum_{q_iq_{i+1} \in q} f_U(\phi(d), q_iq_{i+1}) \tag{11}$$

The MRF model using Eq. (11) is referred to as **VN-MRF**.

3.4. Scope Measure

The remaining problem is how to compute the scope of a document $s(d)$. In this study, we adopt three different approaches – length power, the number of unique terms, and entropy power.

3.4.1. LengthPower. As mentioned in the introduction, according to the scope hypothesis, the document length is affected by the scope: the broader the scope of a document, the longer the document is, when its verbosity is assumed to be fixed. Therefore, the document length could possibly be used as a scope measure according to the scope hypothesis. To derive such a length-based measure, suppose that the scope of a document is a function of document length, i.e., $s(d) = g(|d|)$. Many variants exist for such a function; however, the verbosity and the scope hypotheses help us restrict the possible space for $g(|d|)$, given the following two necessary constraints:

- **SC1**: Scope $g(|d|)$ is a non-decreasing function of $|d|$.
- **SC2**: Verbosity $|d|/g(|d|)$ is a non-decreasing function of $|d|$.

To obtain such a scope measure that would satisfy both SC1 and SC2, we use Heap's law, which is given as follows [Heaps 1978][4]:

$$l_\beta(d) = |d|^\beta$$

where β is an additional constant[5].

The possible range of β is $0 \leq \beta \leq 1$, from SC1 and SC2. Otherwise, $s(d)$ (or $v(d)$) violates SC1 (or SC2) if $\beta < 0$ (or $\beta > 1$). This length-based scope measure $l_\beta(d)$ exactly degenerates into the original unnormalized representation, as a special case, when β = 1 and k = 1, in which case $l_\beta(d)$ = $|d|$, $v(d)$ = 1, and $s(d) = |d|$. The scope measure using $l_\beta(d)$ is called **LengthPower** in this paper.

3.4.2. UniqLength . Another useful scope measure is the number of unique terms $u(d)$, defined as $|\{w|w \in d\}|$. This is reasonable, because a different topic is described using a domain-specific vocabulary or named entities. The more unique terms used in a document, the larger is the scope of the document. The scope measure $u(d)$ is referred to as **UniqLength** in this paper.

3.4.3. EntropyPower . The third scope measure is an entropy-based metric. Previously, the entropy of a document was used to define the homogeneous measure of a document [Bendersky and Kurland 2008], which corresponds to the opposite concept of scope. Another entropy-based metric is the entropy power defined by the exponential of the entropy, which was initially exploited in [Kurland and Lee 2005] to construct the document structure. We compared the entropy with the entropy power in our preliminary experiments and found that the latter outperformed the former because of its similarity to document length or the number of unique terms. Thus, we choose entropy power as our entropy-based metric, and it is defined as follows:

$$h(d) = \begin{cases} \exp\left(-\sum_w p_{ml}(w|d) ln\left(p_{ml}(w|d)\right)\right) & if |d| \geq 1 \\ 0 & otherwise \end{cases}$$

where $p_{ml}(w|d)$ is defined by $c(w,d)/|d|$, which is the maximum likelihood estimation (MLE) of the document language model for d. The scope measure $h(d)$ is called **EntropyPower** in this paper.

4. RETRIEVAL HEURISTICS OF VN RETRIEVAL MODELS

In order to analytically check how differently the VN method satisfies retrieval constraints as compared to the corresponding original model, we present a comparative axiomatic analysis performed under the retrieval constraints introduced by Fang et al. [2004] [6].

4.1. Reference Retrieval Constraints

As in the approach of [Clinchant and Gaussier 2010], we divide the six standard constraints into two different sets – *form constraints* (i.e., TFC1, TFC2, and TDC in [Fang et al. 2004]) and *normalization constraints* (i.e., LNC1, LNC2, and TF-LNC

[4]The Heaps law predicts the number of unique terms in a document from the document length, i.e., the number of unique terms in a corpus increases according to a $k \cdot |d|^\beta$ relationship to the document length. Because the number of unique terms can be used as a scope measure to indicate how broad the topic of the document is, as presented in Section 3.4.2, we use the formula of the Heaps law to approximately predict the number of unique terms using only the document length.

[5]The original form of Heap's law is $\kappa|d|^\beta$, containing the additional parameter κ. Here, we assume that κ is absorbed in k.

[6] Note that our goal in this section is to 'capture' retrieval heuristics of VN retrieval models, but 'not' to refine or improve the standard retrieval constraints of [Fang et al. 2004].

in [Fang et al. 2004]). The form constraints specify the desirable restrictions on the "curve" of a scoring function. Formally, suppose that q consists of a single word w and $f(\phi(d), q)$ is formulated by $g(x, y)$, where x is $c(w, d)$ and y is $idf(w)$. Then, TFC1, TFC2, and TDC [Clinchant and Gaussier 2010; Fang et al. 2011] correspond to, respectively:

$$\frac{\partial g(x,y)}{\partial x} > 0,\ \frac{\partial^2 g(x,y)}{\partial^2 x} < 0,\ \frac{\partial g(x,y)}{\partial y} > 0$$

It can be easily shown that TFCs and TDC are satisfied for all three normalized functions. This is a natural result, because our normalization only linearly transforms the term frequency and retains the original model, without any change to the basic concepts of the original model.

The normalization constraints describe the necessary properties of a retrieval model for the case in which document-specific quantities such as length, verbosity, and scope are different across documents. According to [Fang et al. 2011], each normalization constraint can be equivalently described by how the score of a document changes after applying a perturbation operator to the document. We introduce three perturbation operators called **PAN**, **PLS**, and **PAR** that correspond to LNC1, LNC2, and TF-LNC, respectively, as follows[7]:

1) **PAN** (**P**erturbation of **A**dding **N**oise Words): PAN is an operator for adding noise terms, denoted by ψ_{AN}. Given d, $\psi_{AN}(d)$ is obtained by adding K noise words $v_1 \cdots v_K$ to d, i.e., $\psi_{AN}(d) = d\, v_1 \cdots v_K$, where $v_i \notin q$. When $d_2 = \psi_{AN}(d_1)$, $|d_2| = |d_1| + K$ and $c(w, d_2) = c(w, d_1)$ for all $w \notin q$.

2) **PLS** (**P**erturbation of **L**ength **S**caling): PLS is a length scaling operator, denoted by ψ_{LS}. Given d, $\psi_{LS}(d)$ is obtained by concatenating **all query words** in d K times and by scaling the length of d up to K times. When $d_1 = \psi_{LS}(d_2)$, $|d_1| = K \cdot |d_2|$ and $c(w, d_1) = K \cdot c(w, d_2)$ for all $w \in q$. Note that the concatenation is only applied to query words, not necessarily to non-query words. The non-query words in d might (or might not) be kept in $\psi_{LS}(d)$. In the extreme case, all the non-query words do not appear in $\psi_{LS}(d)$, being replaced with other non-query words.

3) **PAR** (**P**erturbation of **A**dding **R**elevant Words): PAR is an operator for adding a single relevant word, denoted by ψ_{AR}. Given d, $\psi_{AR}(d)$ is obtained by appending a single query word $w \in q$, i.e., $\psi_{AR}(d) = d\, w \cdots w$ (i.e., the attached number of w is K) K times. When $d_1 = \psi_{AR}(d_2)$, $|d_1| = |d_2| + K$, $c(w, d_1) = c(w, d_2) + K$ for a given single word $w \in q$, and $c(w', d_1) = c(w', d_2)$ for all $w' \neq w$.

Here, K is a perturbation parameter. LNC1, LNC2, and TF-LNC can now be equivalently described as follows:

LNC1: If $d_2 = \psi_{AN}(d_1)$, $f(d_1, q) \geq f(d_2, q)$ for $K \geq 1$.

LNC2: If $d_1 = \psi_{LS}(d_2)$, $f(d_1, q) \geq f(d_2, q)$ for $K > 1$.

TF-LNC: Let $q = \{w\}$ be a query with only one term w. If $d_1 = \psi_{AR}(d_2)$, $f(d_1, q) > f(d_2, q)$ for $K \geq 1$.

The perturbation operator PLS for LNC2 is slightly different from the original version of LNC2 [Fang et al. 2004; Fang et al. 2011]. In the original version, d_1 is fully copied to d_2, making them identical. In our PLS, only query words are concatenated K times to d_1, and no further assumption is made about non-query words. Therefore, PLS is the generalized version of the original operator, including the original version as a special case. This generalization does not cause any inconsistency in the known analysis results of LNC2; the analysis results reported in [Fang et al. 2004;

[7] PAN, PLS, and PAR correspond to TN, LV3, and TG1, respectively, in [Fang et al. 2011]

Table III. Percentages of A_1 being satisfied using all non-stopwords in all queries from three collections (ROBUST, WT10G, and GOV2) and three query types (sk, sv, and lv). The columns A_1^{Okapi} and A_1^{DP} indicate the conditions $df(w) \leq N/2$ and $c(w,d) \geq |d|p(w|C)$, respectively.

	ROBUST		WT10G		GOV2	
	A_1^{Okapi}	A_1^{DP}	A_1^{Okapi}	A_1^{DP}	A_1^{Okapi}	A_1^{DP}
sk	99.5%	99.1%	98.5%	96.2%	99.3%	95.6%
sv	99.4%	98.3%	99.1%	95.1%	98.9%	93.5%
lv	99.3%	98.3%	99.3%	95.1%	99.2%	92.9%

Fang et al. 2011] for LNC2 are also still consistently accepted with our PLS operator. To see the difference more clearly, Algorithm 1 summarizes the detailed description of our PLS operator.

Algorithm 1 The detailed procedure of PLS

Step 1) Given d, we apply the original PLS operator of [Fang et al. 2004]' to d to obtain d'; d' is obtained by simply concatenating all words in d K times
Step 2) Given d', $\psi_{LS}(d)$ is obtained after applying the following procedure:
Let d' be $w_1 \cdots w_{|d'|}$
Initialize $\psi_{LS}(d)$ as an empty document.
for $i \leftarrow 1, |d|'$ **do**
  if $w_i \in q$ **then**
    $\psi_{LS}(d) \leftarrow \psi_{LS}(d)\ w_i$
  else
    $\psi_{LS}(d) \leftarrow \psi_{LS}(d)\ w'$ ($w' \notin q$) where w' is randomly chosen from $\mathcal{V} \setminus q$.
  end if
end for

4.2. Analysis Results of Normalization Constraints

4.2.1. Assumption. Before presenting our analysis results of the three normalization constraints, we make the following assumption:

— A_1: For any query word $w \in q$, w is assumed to be a content-bearing word (i.e., $df(w) \leq N/2$, and $c(w,d) \geq |d|p(w|C)$ for any document d in the collection).

Empirically, A_1 holds well in usual cases when we filter out stopwords. Table III lists the percentage of A_1 being satisfied using all non-stopwords in all queries from three different collections and three query types. (Refer to Section 5.1 for a description of the collections and query types.) As shown in Table III, $c(w,d) \geq |d|p(w|C)$ is satisfied in more than 98% of the documents for all query words in ROBUST, more than 95% in WT10G, and more than about 93% in GOV2. The condition $df(w) \leq N/2$ is satisfied for more than 98% of the query terms.

4.2.2. Analysis Results. There exist necessary conditions common for all VN retrieval models under A_1 to be satisfied for each normalization constraint. Table IV summarizes the analysis results of the general and the special cases of scope using Length-Power and UniqLength for VN retrieval models, relative to the original models [8].

[8]Note that the analysis results are obtained from DP and Okapi, not from MRF. For MRF, we do not separately carry out axiomatic analysis, since it is not a base model like DP and Okapi, but being an extension of a base model (i.e. the scoring function of MRF is defined in terms of the main function of its base model). Thus, it is reasonable to assume that the normalization heuristics of MRF will not be significantly different from its base model, without separate analysis.

Table IV. Analysis results of the original and VN retrieval models for three normalization constraints – LNC1, LNC2, and TF-LNC – under A_1.

	LNC1	LNC2	TF-LNC
Original [Fang et al. 2004]	Yes	Yes	Yes
Verbosity-normalized (General)	C_1	C_2	C_3
Verbosity-normalized (LengthPower)	Yes	Yes	Yes
Verbosity-normalized (UniqLength)	C_1	C_2	Yes
Verbosity-normalized (EntropyPower)	C_1	C_2	C_4

Table V. Percentages of C_4 being satisfied using all non-stopwords in all queries from three collections (ROBUST, WT10G, and GOV2) and three query types (sk, sv, and lv).

	ROBUST	WT10G	GOV2
sk	99.99%	99.97%	99.98%
sv	99.99%	99.98%	99.98%
lv	99.99%	99.98%	99.96%

Table IV uses the notations introduced by [Fang et al. 2004], where "Yes" and "C_x" indicate that the corresponding model satisfies the particular constraint in the absence of conditions and under particular conditions, respectively. The specific conditions are

- C_1: $v(d_2) \geq v(d_1)$
- C_2: $s(d_1) \geq s(d_2)$
- C_3: $K/c(w, d_2) \geq v(d_1)/v(d_2) - 1$
- C_4: $s(d_2) \leq (|d_2|/c(w, d_2))^2$

where C_1, C_3 and C_4 are sufficient but not necessary conditions to satisfy the particular constraint. Some derivations of the conditions are given in Appendix C-E.

As shown in Table IV, an original method satisfies all three constraints unconditionally under A_1 according to [Fang et al. 2004], whereas a VN method requires additional conditions that depend on the choice of scope measure. An exceptional case is LengthPower, in which all constraints are satisfied unconditionally.

Among the three constraints, TF-LNC is satisfied under LengthPower and UniqLength, the detailed proofs of which are presented in Appendix D. Under EntropyPower, TF-LNC is satisfied for almost all query words in our test collection, as shown in Table V; C_4 is satisfied in more than 99.9% of the documents for all query words in ROBUST, WT10G and GOV2. Therefore, we do not explore TF-LNC further in this paper.

4.3. Normalization Heuristics of VN Retrieval Models (Case: UniqLength and EntroyPower)

In this section, we discuss the retrieval behaviors entailed from the VN method in the cases of UniqLength and EntropyPower, with respect to the original method. In our discussion, PAN and PLS are further divided into two different types – V-type and S-type – which refer to *verbosity-increasing* and *scope-broadening* perturbations, respectively. The definitions of these types of operators are as follows:

(1) **V-type perturbation**: The operator $\psi(\cdot)$ is called *V-type* if the perturbation does not *increase* the scope of the document, i.e., if $d_1 = \psi(d_2)$ and ψ is V-type, $s(d_1) \leq s(d_2)$.
(2) **S-type perturbation**: The operator $\psi(\cdot)$ is called *S-type* if the perturbation does not *decrease* the scope of the document, i.e., if $d_1 = \psi(d_2)$ and ψ is S-type, $s(d_1) \geq s(d_2)$

We then reexamine how the original and VN models satisfy LNCs on V-type and S-type PAN and PLS.

The notable result is that C_1 and C_2 correspond to a relaxed penalization of a scope-broadened document, and a strict penalization of a verbosity-increased document, respectively.

First, we present the first heuristic H1 and discuss its derivation from C_1:

4.3.1. ***H1: Relaxed penalization of scope-broadened documents****. The VN retrieval method performs a relaxed penalization of a scope-broadened document after performing PAN. (from LNC1)*

To derive H1, we divide PAN into V-PAN and S-PAN. **V-PAN** denotes *verbosity-increasing* PAN, where K added noise words are covered by the original scope of the document. **S-PAN** denotes the *scope-broadening* PAN, where K added noise words describe *new* contents that are not covered by the scope of the original document. In terms of V-type and S-type, V-PAN and S-PAN can be defined as follows:

(1) **V-PAN**: V-PAN is a specific type of PAN, being V-type.
(2) **S-PAN**: S-PAN is a specific type of PAN, being S-type.

Suppose that d_1 and $d_2 = \psi_{AN}(d_1)$ are the given documents for LNC1. Then, we can show that the VN model often does *not* penalize d_2 for S-PAN; instead, it penalizes d_2 for V-PAN. On the other hand, the original model *always* penalizes d_2 for both S-PAN and V-PAN. Thus, the VN model imposes a type of relaxed penalization to *scope-broadened* documents after PAN, with respect to the original model.

Equivalently, the heuristic H1 can be rewritten in the form of a retrieval constraint as follows:

H1-LNC: If $d_2 = \psi_{AN}(d_1)$ and ψ_{AN} is S-PAN, $f(d_1, q) \leq f(d_2, q)$ for $K \geq 1$ with the following condition C_5 and C_6, for VN-Okapi and VN-DP, respectively:

- C_5:

$$\frac{v(d_1) - v(d_2)}{K} \geq \frac{b}{1-b}\frac{1}{avgs} \tag{12}$$

- C_6:

$$\frac{v(d_1) - v(d_2)}{K} \geq \frac{1}{s(d_1)}\frac{p(w|C) + p_{ml}(w|d)s(d_1)\mu^{-1}}{p_{ml}(w|d) - p(w|C)} \tag{13}$$

where $p(w|d) > p(w|C)$ is additionally assumed in C_6[9].

Compared to LNC1, the consequence part of H1-LNC conditionally entails the negation of LNC1, implying that the VN model often *prefers* some of the scope-broadened documents resulting from S-PAN, although the original model does not [10].

[9]For C_6, assuming an extreme case where the query is a very highly topical, i.e., $p_{ml}(w|d) \gg p(w|C)$ or $r \to \infty$), C_6 is simplified as:

$$\frac{v(d_1) - v(d_2)}{K} \geq \frac{1}{\mu} \tag{14}$$

[10]From Eq. (12) and Eq. (13), when $(v(d_1) - v(d_2))/K$ is sufficiently large, C_5 (or C_6) can be satisfied both for VN-Okapi and VN-DP. This case can appear if $v(d_1)$ is large and $s(d_1)$ is small, but it is not always

• *Example of H1*: Here, we present examples of S-PAN and V-PAN. Suppose that we use UniqLength as the scope measure and a document consisting of passages that are disjoint in scope. Formally, let $\boldsymbol{g}$, h, and x be passages, and assume that $\boldsymbol{g}$, h, and x have no common or overlapping content, where $\boldsymbol{g}$ denotes a relevant passage and h and x are non-relevant, i.e., $c(w,\boldsymbol{g}) > 0$, $c(w,h) = c(w,x) = 0$ for query word $w \in q$. Examples of S-PAN and V-PAN are as follows:

Example of S-PAN:

$d_1 = \boldsymbol{g}\ \boldsymbol{g}\ h\ h$
$d_2 = \boldsymbol{g}\ \boldsymbol{g}\ h\ h\ x$

Example of V-PAN:

$d_1 = \boldsymbol{g}\ \boldsymbol{g}\ h\ h$
$d_2 = \boldsymbol{g}\ \boldsymbol{g}\ h\ h\ h$

For both examples, the query relevant content is not changed after PAN. Because these two examples are PAN examples, the original method always prefers d_1 to d_2, irrespective of the PAN type. However, C_1 is satisfied only for V-PAN, because $s(d_2) = s(d_1)$ and $v(d_2) \geq v(d_1)$, and not clearly for S-PAN, because $v(d_2) < v(d_1)$ is plausible due to $s(d_2) > s(d_1)$. Therefore, the VN method prefers d_2 in V-PAN and not always in S-PAN.

• *Derivation of H1*: To show how the VN model behaves differently toward V-PAN and S-PAN, we first rewrite C_1 by $s(d_2) - s(d_1) \leq K/v(d_1)$, implying that the scope of the new document d_2 must not be increased considerably after performing PAN.

i) Case: V-PAN

First, V-PAN does not increase $s(d_2)$ according to the definition of a V-type perturbation, resulting in $s(d_2) - s(d_1) \leq 0$. As a result, it is clear that C_1 is always true for V-PAN, finally making LNC1 true. Thus, for V-PAN, there is no difference between the original and the VN models in satisfying LNC1.

For example, suppose that we use UniqLength as the scope measure and consider a given V-PAN in which all K words already occur in d_1. In this case, $s(d_2) = s(d_1)$ because no new words occur in d_2. Thus, C_1 is equivalent to $s(d_2) - s(d_1) = 0 \leq K/v(d_1)$, which is true irrespective of K.

ii) Case: S-PAN

Second, S-PAN increases the scope after performing PAN according to the definition of an S-type perturbation, resulting in $s(d_2) - s(d_1) \geq 0$. Therefore, C_1 is not always true.

For example, suppose that we use UniqLength as the scope measure again, and consider a given S-PAN in which all K words are new and different from each other. In this case, $s(d_2) = s(d_1) + K$, and C_1 is equivalent to $K \leq K/v(d_1)$; however, C_1 is usually not satisfied because $v(d_1) \geq 1$.

Instead, it often satisfies the *negation* of LNC1. Consider the same S-PAN example in which all K words are different from each other and assume that we use VN-DP as an example retrieval model. C_6 is then equivalent to:

$$\frac{v(d_1) - 1}{1 + K/s(d_1)} \geq \frac{p(w|C) + p_{ml}(w|d)s(d_1)\mu^{-1}}{p_{ml}(w|d) - p(w|C)} \tag{15}$$

There exist a number of situations in which C_6 is true according to Eq. (15) (i.e., $v(d_1)$ is sufficiently large, or K added words are highly topical ($p(w|d) \gg p(w|C)$) and μ

true. Otherwise, C_5 (or C_6) can be satisfied, according to the choice of a retrieval parameter value or a term discrimination value of a query word; for VN-Okapi, C_5 is satisfied if b is sufficiently large; for VN-DP, C_6 is satisfied if μ is sufficiently large and the query word is highly topical.

is reasonably large). Therefore, for S-PAN, the VN model often does not satisfy LNC1, in contrast to the original model that always satisfies LNC1 □.

Next, we present the heuristic H2 and discuss its derivation from C_2:

4.3.2. ***H2: Strict penalization of verbosity-increased documents****. The VN retrieval method imposes a strict penalization of a verbosity-increased document after performing PLS (from LNC2).*

As was performed on PAN, we divide PLS into V-PLS and S-PLS. **V-PLS** denotes *verbosity-increasing* PLS, where the non-query words after PLS is performed are covered by the original scope of the document, thereby increasing verbosity. **S-PLS** denotes the *scope-broadening* PLS, where the non-query words after PLS is performed introduce *new* contents that are not covered by the scope of the original document. In terms of V-type and S-type, V-PLS and S-PLS can be defined as follows:

(1) **V-PLS**: V-PLS is a specific type of PLS, being V-type.
(2) **S-PLS**: S-PLS is a specific type of PLS, being S-type.

Given two documents $d_1 = \psi_{LS}(d_2)$ and d_2 for LNC2, from the definition of C_2 (i.e., $s(d_1) \geq s(d_2)$), LNC2 is satisfied only if the scope of the original document increases after PLS is performed. Therefore, the VN model *prefers* (or does not penalize) only d_1 for S-PLS because it increases the scope; instead, it *penalizes* d_1 for V-PLS, which decreases the scope. As such, the VN model imposes a strict penalization of a verbosity-increased document after PLS.

Equivalently, the heuristic H2 can be rewritten in a form of retrieval constraint as follows:

H2-LNC: If $d_1 = \psi_{LS}(d_2)$ and ψ_{LS} is V-PLS, $f(d_1, q) \leq f(d_2, q)$ for $K > 1$.

Compared to LNC2, the consequence part of H2-LNC is the negation of that of LNC2, implying that the VN model *always* penalizes verbosity-increased documents resulting from S-PLS, although the original model does *not* (i.e, prefers them).

• *Example of H2*: We present examples of S-PLS and V-PLS. Suppose that we use UniqLength as the scope measure and a document consisting of passages that are disjoint in scope. Formally, let $\boldsymbol{g}$, h, x, and y_i be passages with equal length (i.e., $|\boldsymbol{g}| = |h| = |x| = |y_i|$) and unit scope (i.e., $s(g) = s(h) = s(x) = s(y_i) = 1$), and assume that $\boldsymbol{g}$, h, x, and y_i have no common or overlapping content, where $\boldsymbol{g}$ denotes a relevant passage and h, x, and y_i are non-relevant, i.e., $c(w, \boldsymbol{g}) > 0$, $c(w, h) = c(w, x) = c(w, y_i) = 0$ for query word $w \in q$. Examples of S-PLS and V-PLS are as follows:

Example of S-PLS:

$d_1 = \boldsymbol{g}\ \boldsymbol{g}\ h\ x\ y_1\ y_2$
$d_2 = \boldsymbol{g}\ h\ x$

Example of V-PLS:

$d_1 = \boldsymbol{g}\ \boldsymbol{g}\ h\ h\ h\ h$
$d_2 = \boldsymbol{g}\ h\ x$

For both examples, $|d_2| = 2|d_1|$, $c(w, d_1) = 2c(w, d_2)$ for $w \in q$, i.e., the query-relevant content is copied twice. The example of S-PLS introduces two *new* non-relevant passages y_1 and y_2 that are not given in d_2, whereas the example of V-PLS does not introduce any new passage but only repeats the previously mentioned non-relevant passage h.

Because these two examples are PLS examples, the original method always prefers d_1 to d_2, irrespective of the PLS type. However, C_2 is satisfied only for S-PLS and not for V-PLS, from $s(d_1) = 5 > s(d_2) = 3$ in S-PLS and $s(d_1) = 2 < s(d_2) = 3$ in V-PLS. Therefore, the VN method prefers d_1 only in S-PLS and not in V-PLS.

Table VI. The summary of the normalization behaviors of the original and VN models for four perturbations – S-PAN, V-PAN, S-PLS, and V-PLS – under A_1 in which d is original document, and $\psi_{AN}(d)$ (or $\psi_{LS}(d)$) is the perturbed documents of d after PAN (or PLS).

ψ	Verbosity-normalized model (UniqLength, EntropyPower)	Original model
S-PAN	$f(d,q) \leq f(\psi_{AN}(d),q)$ if C_5(or C_6) is true $f(d,q) \geq f(\psi_{AN}(d),q)$ Otherwise	$f(d,q) \geq f(\psi_{AN}(d),q)$
V-PAN	$f(d,q) \geq f(\psi_{AN}(d),q)$	
S-PLS	$f(d,q) \leq f(\psi_{LS}(d),q)$	$f(d,q) \leq f(\psi_{LS}(d),q)$
V-PLS	$f(d,q) \geq f(\psi_{LS}(d),q)$	

• *Derivation of H2*: From the definitions of V-type and S-type perturbations, it is trivial to show that C_2 is true for V-PLS but false for S-PLS. Therefore, for S-PLS, the VN model does not satisfy LNC2, in contrast to the original model which always satisfies LNC2 □.

4.3.3. Summary. Table VI summaries the normalization behaviors of the original and VN models in response to S-PAN, V-PAN, S-PLS, and V-PLS.

For V-PAN and S-PLS, the VN model leads to the same normalization heuristics as those of the original model. For S-PAN and V-PLS, however, the normalization behaviors are completely different between the original and the VN models; for S-PAN, the VN model often does *not* penalize the new document, whereas the original model *always* penalizes it; for V-PLS, the VN model *always* penalizes the new document, whereas the original model does *not* penalize it.

Overall, the normalization heuristics entailed from the VN model are dependent on whether a perturbation is V-type or S-type. For a V-type perturbation, the VN model imposes a strict penalization of a verbosity-increased document (i.e., entailing H1), irrespective of PAN or PLS. For an S-type perturbation, the VN model unlikely penalizes a scope-broadened document (i.e., entailing H2). On the other hand, the normalization heuristics of the original model are dependent on whether a perturbation is PAN or PLS, not on whether it is V-type or S-type. For PAN, the original model penalizes a new document, irrespective of whether the document is verbosity-increased or scope-broadened. For PLS, it does not penalize the new document.

5. EXPERIMENTAL SETTING

5.1. Experimental Setup

For evaluation, we used three different standard TREC collections – ROBUST, WT10G, and GOV2. Table VII lists the basic statistics of each test collection, where *NumDocs* is the number of documents, *NumWords* is the total number of word occurrences in each collection, *TopicSet* is the range of topic numbers used for training and testing, and *Avg* of $|d|$, $h(d)$, and $v(d)$ indicates the average length, entropy power, and verbosity[11], respectively, in a given collection. CoeffVar is the corresponding *coefficient of variance*, which is defined as the ratio of the standard deviation to the mean. The interesting statistic is CoeffVar of $v(d)$, which indicates the differences among the verbosities of documents in a collection. ROBUST has the most similar verbosities across documents, whereas GOV2 has the most different verbosities. This is because many documents in ROBUST are newspaper documents, for example, from Financial Times and Los Angeles Times, which are more homogeneous collections. In contrast, the web documents in GOV2 are more heterogeneous.

All experiments were performed using the Lemur toolkit (version 4.12). We carried out standard preprocessing by applying the Porter stemmer and removing stopwords

[11]Entropy power is used as scope measure for $v(d)$.

Table VII. Statistics of each test collection. CoeffVar denotes the coefficient of variation.

Statistics	ROBUST	WT10G	GOV2
NumDocs	528,156	1,692,096	25,205,179
NumWords	572,180	6,346,858	40,002,579
TopicSet	Q301−450 Q601−700	Q451−550	Q701−850
Avg of $\|d\|$ (CoeffVar)	233.34 (2.39)	400.25 (6.06)	690.8 (2.86)
Avg of $h(d)$ (CoeffVar)	107.77 (0.81)	109.60 (1.45)	109.85 (0.98)
Avg of $v(d)$ (CoeffVar)	1.77 (0.91)	2.95 (5.51)	6.11 (7.17)

from the standard INQUERY stoplist [Allan et al. 2000]. To cover different types of queries, we follow the setting used in [Zhai and Lafferty 2001], where four combinations are used: short keywords (**sk**, title), short verbose (**sv**, description), long keywords (**lk**, concept), and long verbose (**lv**, title, description, and narrative). In our test topic sets, because lk is not available, the other three types were examined. We use MAP (mean average precision) and P@5 (precision at top 5 documents) as the evaluation measures [Croft et al. 2009].

For each query, our evaluation is based on the top 1,000 documents retrieved. We also report significance test results by a non-directional paired t-test at 0.95 confidence level. For the significance test, we use all *per-topic* performances in a collection, i.e., the number of performance difference samples used for the t-test is the same as the total number of topics in a given collection [12].

5.2. Parameter Tuning

Several tuning parameters are present in the retrieval methods – DP: μ and Okapi: b, k_1, and k_3. Given a test topic set consisting of 50 queries, each parameter was tuned using the other topic sets in the same test collection as the development set[13]. The search space for each parameter is given as follows:

• μ: { 100, 200, 300, 400, 500, 600, 800, 1000, 1500, 2000, 2500, 3000, 4000, 5000, 7000, 10000, 15000, 20000 }

• b: {0, 0.001, 0.003, 0.005, 0.007, 0.01, 0.02, 0.03, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}

[12] In Section 5.2, we introduce a K-fold cross-validation to avoid the optimization of the retrieval parameters to the test set. However, note that we do not use *per-fold* performances to perform the significance test but simply use all per-topic performances. To the best of our knowledge, this type of significance test is an IR-specific setting that is different from the other types of significance test used in non-IR literatures.

[13]Here, a topic set consists of 50 queries, which was were created in each year of TREC. For example, in ROBUST, as 250 queries are available, there are 5 topic sets, namely, TREC6(Q301−Q350), TREC7(Q351−Q400), TREC8(Q401−Q450), ROBSUT03(Q601−Q650), and ROBUST04(Q651−Q700). Parameters used when testing 50 queries in each topic set are trained using the other 200 queries in other topic sets as training data. In other words, for testing 50 queries in TREC6, queries Q351−Q450 and Q601−Q700 in TREC7, TREC8, ROBUST03, and ROBUST04 are used as training data. For testing queries in TREC7, queries in TREC6, TREC8, ROBUST03, and ROBUST04 are used as the training set, and so on. Therefore, for ROBUST, we use a five-fold cross validation for parameter tuning, whose folds are fixed. For WT10G, where 100 queries are used, we have two topic sets, namely TREC9(Q451−Q500) and TREC10(Q501−Q550). For testing 50 queries in TREC9, we use queries in TREC10 as the training set, and vice versa. Thus, for WT10G, we use a two-fold cross validation for parameter tuning. Similarly, for GOV2, 150 queries are available, so we have three topic sets, namely TREC2004(Q701−Q750), TREC2005(Q751−Q800) and TREC2006(Q801−Q850). For testing 50 queries in TREC2004, queries in TREC2005 and TREC2005 are used as the training data. Thus, for GOV2, we use a three-fold cross validation for parameter tuning.

Table VIII. MAP performance comparison of DP and VN-DP on three collections ROBUST, WT10G, and GOV2; three different query types sk, sv, and lv; and three different scope measures LengthPower (β), UniqLength, and EntropyPower. The row titled "baseline" indicates the original model. The symbols * indicate that a run of the VN method shows statistically significant improvement over the baseline in the t-test, at 0.95 confidence level.

	Method	DP (or VN-DP)		
		ROBUST	WT10G	GOV2
sk	baseline	0.2447	0.1963	0.2907
	LengthPower(0.25)	0.2252	0.1649	0.2403
	LengthPower(0.5)	0.2401	0.1953	0.2823
	LengthPower(0.75)	0.2457	0.1968	0.2930
	LengthPower(0.9)	0.2460*	0.1963	0.2913
	UniqLength	0.2472*	0.2046	0.3055*
	EntropyPower	**0.2481***	**0.2120***	**0.3099***
sv	baseline	0.2260	0.1909	0.2455
	LengthPower(0.25)	0.2312	0.1790	0.2350
	LengthPower(0.5)	**0.2443***	0.2103*	0.2633*
	LengthPower(0.75)	0.2396*	0.2044*	0.2569*
	LengthPower(0.9)	0.2319*	0.1946*	0.2487*
	UniqLength	0.2385*	0.2109*	0.2671*
	EntropyPower	0.2440*	**0.2196***	**0.2826***
lv	baseline	0.2707	0.2469	0.2864
	LengthPower(0.25)	0.2697	0.2249	0.3060*
	LengthPower(0.5)	0.2765*	0.2506	0.3133*
	LengthPower(0.75)	0.2762*	0.2532*	0.3005*
	LengthPower(0.9)	0.2725*	0.2501*	0.2914*
	UniqLength	0.2759*	0.2553*	0.3083*
	EntropyPower	**0.2799***	**0.2614***	**0.3248***

- k_1: {0.25, 0.3, 0.4, 0.5, 0.6, 0.8, 1.0, 1.2, 1.5, 1.8, 2.0, 2.5, 3.0}
- k_3: fixed at 1,000

In our preliminary experiments, we found that LengthPower for $s(d)$ can suffer from the parameter scaling problem, in which the optimal parameter ranges of μ and k_1 in the VN methods differ from the known ranges of the original model. For instance, when $\beta = 0.25$, it was found that μ was optimal at a value of less than 100, which is beyond the normal parameter range. To resolve the scaling problem, we substitute $avgv$ for k, instead of setting k to 1, such that $c(w, \phi(d))$ would become $c(w, d)$ on average. This consideration leads to the following parameter scaling:

$$k_1 \leftarrow k_1 \cdot avgv^{-1},\ \mu \leftarrow \mu \cdot avgv^{-1}$$

This parameter scaling is applied only to LengthPower, and not to the others. No such parameter scaling problem occurs in the case of UniqLength and EntropyPower.

6. EXPERIMENTAL RESULTS

This section reports the comparative results of the original and the VN retrieval method for Okapi, DP and MRF.

6.1. DP vs. VN-DP

Table VIII show the comparative results (MAP) of DP and VN-DP under three different scope measures, UniqLength, EntropyPower, and LengthPower(β), which are denoted as $l_\beta(d)$, $u(d)$, and $h(d)$, respectively.

Generally, it is observed that VN-DP improves original DP. These improvements are statistically significant for almost all test collections and all query types (for both

Table IX. P@5 performance comparison of DP and VN-DP for three collections – ROBUST, WT10G, and GOV2.

	Method	DP (or VN-DP)		
		ROBUST	WT10G	GOV2
sk	baseline	0.4924	0.3120	0.5678
	LengthPower(0.5)	0.4707	0.3360	0.5409
	LengthPower(0.75)	0.4851	0.3220	0.5611
	LengthPower(0.9)	0.4924	0.3080	0.5691
	UniqLength	0.4956	0.3620*	0.5906
	EntropyPower	**0.4972**	**0.3640***	**0.6416***
sv	baseline	0.4466	0.3880	0.5208
	LengthPower(0.5)	0.4811*	0.4000	0.5383
	LengthPower(0.75)	0.4699*	0.3960	0.5409
	LengthPower(0.9)	0.4530	0.3820	0.5275
	UniqLength	0.4755*	0.4060	0.5826*
	EntropyPower	**0.4932***	**0.4300***	**0.6309***
lv	baseline	0.5414	0.4460	0.6228
	LengthPower(0.5)	0.5518	0.4660	0.6188
	LengthPower(0.75)	0.5510	0.4560	0.6295
	LengthPower(0.9)	0.5526*	0.4520	0.6282
	UniqLength	0.5542*	0.4560	0.6456*
	EntropyPower	**0.5631***	**0.4700**	**0.6644***

UniqLength and EntropyPower), often resulting in an improvement of 10%. The improvement tends to be larger on Web collections (i.e., WT10G and GOV2) than for ROBUST. A possible reason is that the Web collections have higher CoeffVar of $v(d)$ because of the heterogeneity of documents, and thus, they could gain more from our verbosity normalization.

Among the three scope measures, EntropyPower is the best, and it outperforms UniqLength and LengthPower for most topic sets. UniqLength is slightly better than LengthPower; however, the difference in their performances is not significant. When the best β value is adopted for each query type, LengthPower can often show performance similar to that of UniqLength.

Interestingly, VN-DP leads to significant improvements, more in verbose queries (i.e., sv and lv) than in keyword queries. For EntropyPower, VP-DP causes an improvement of 1.55% in ROBUST for short keyword queries, 8.2% in WT10G, and 6.64% in GOV2. The corresponding improvements are much larger for short verbose queries, being 8.05% in ROBUST, 16.66% in WT10G, and 15.11% in GOV2, and they are also large for long verbose queries. Restricting our discussion to DP, these results strongly support that the use of heuristics H1 and H2 is indeed important, especially for verbose queries.

In addition, Table IX shows the performances of P@5 for VN-DP, as compared to that of DP, based on the MAP-optimized free-parameters' values used in Table VIII. One reason for using the same retrieval parameters instead of directly optimizing P@5 is that the concavity of the performance curve was smoother in MAP than in P@5, thereby avoiding the use of far-from-optimal parameter values. As similarly mentioned by [Kurland and Lee 2009], this choice helps us to examine whether the improved performance in MAP causes severe degradation or significant improvement in the precision, which is an often important metric in some IR applications such as Web search.

Under EntropyPower, the improvement of P@5 from DP to VN-DP is significant in most cases, often being larger than that of MAP for short keyword and verbose queries. VP-DP improves over DP by about 16.67% on WT10G and about 13.00% on GOV2 for short keyword queries, and 13.00% on WT10G and 21.14% on GOV2 for short verbose queries. Exceptional cases are found for keyword queries on ROBUST and for long verbose queries on WT10G, where the improvement in the precision is not statistically

Table X. MAP performance comparison of Okapi and VN-Okapi for three collections, ROBUST, WT10G, and GOV2. The symbols * indicate that a run of the VN method shows statistically significant improvement over the baseline in the t-test at 0.95 confidence level.

	Method	Okapi (or VN-Okapi)		
		ROBUST	WT10G	GOV2
sk	baseline	0.2444	0.1946	0.2920
	LengthPower(0.5)	0.2451	0.1957	0.2897
	LengthPower(0.75)	0.2454	0.1994*	0.2923
	LengthPower(0.9)	0.2452	0.1944	0.2923
	UniqLength	**0.2483***	0.1997	**0.3035***
	EntropyPower	0.2477*	**0.2071***	0.3004*
sv	baseline	0.2247	0.1853	0.2498
	LengthPower(0.5)	0.2279*	0.1806	0.2527
	LengthPower(0.75)	0.2263	0.1872	0.2530*
	LengthPower(0.9)	0.2271*	0.1878	0.2529*
	UniqLength	0.2267	0.1936	**0.2607***
	EntropyPower	**0.2303***	**0.1968***	0.2599*
lv	baseline	0.2619	0.2344	0.3012
	LengthPower(0.5)	0.2647*	0.2307	0.3022
	LengthPower(0.75)	0.2640*	0.2341	0.3009
	LengthPower(0.9)	0.2631	0.2366	0.3018
	UniqLength	**0.2663***	0.2368*	0.3063*
	EntropyPower	0.2659*	**0.2415***	**0.3074***

significant. Therefore, at least for EntropyPower, the results imply that the significant improvements of MAP using VN-DP are caused by the increased performance in P@5. For UniqLength, however, the impact of P@5 and its contribution to MAP is less clear than that for EntropyPower, although some noticeable improvements are observed. For LengthPower, most improvements of P@5 are not statistically significant. This implies that performance metrics other than P@5, such as recall, might be the major factors causing the significant improvement in MAP for LengthPower.

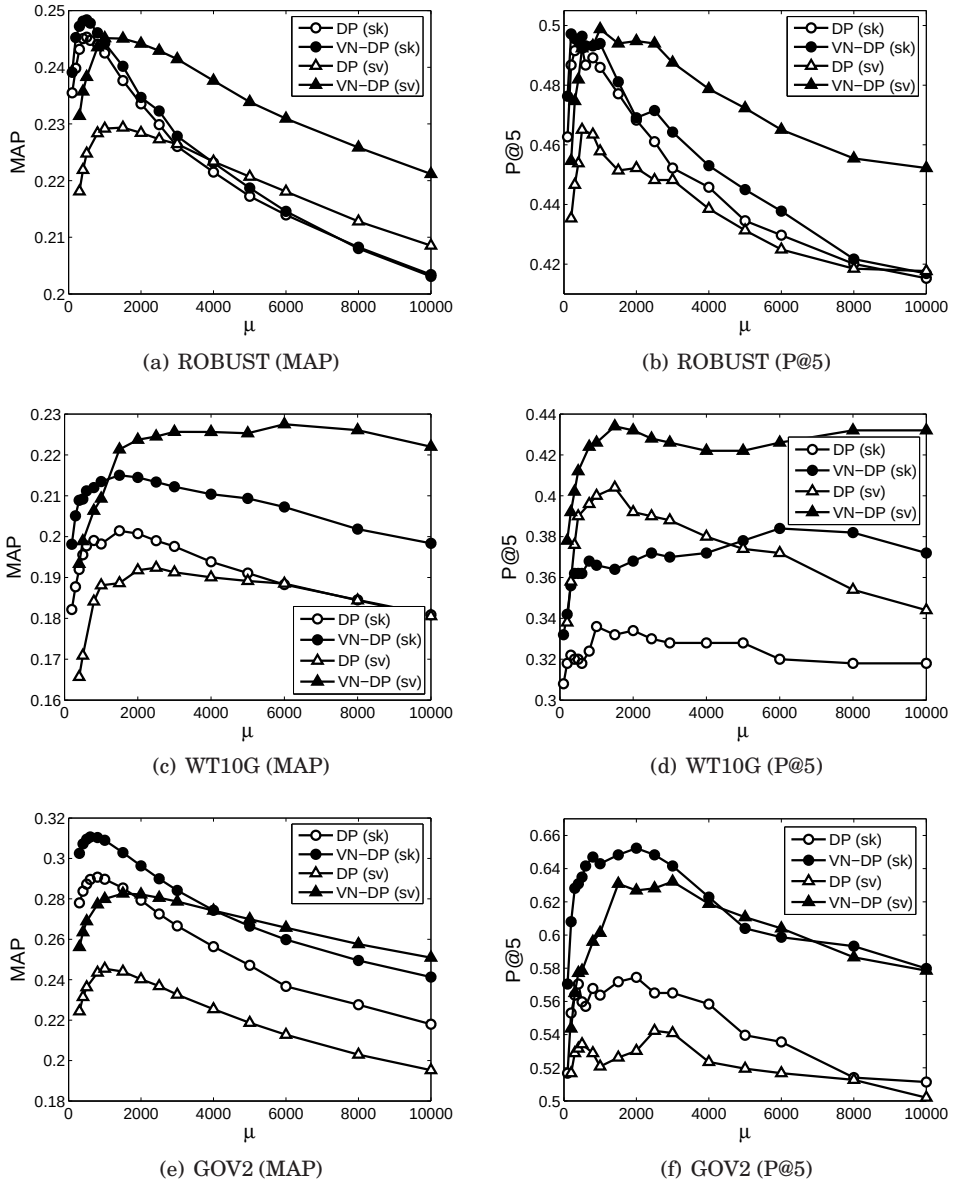
For further comparison, Figure 1 shows the performance curves of the original DP and VN-DP using EntropyPower, plotted by varying μ for MAP and P@5. For both measures, VN-DP is always better than DP, for almost all μ values and in all test collections and query types. The shapes of the curves of DP and VN-DP are similar, and the optimal ranges of μ are also fairly similar. This similarity between DP and VN-DP is also observed in the case of UniqLength, although we do not present the curves for this scope measure here, in the interest of conciseness.

6.2. Okapi vs. VN-Okapi

Table X show the comparative results (MAP) of Okapi and VN-Okapi under three different scope measures – UniqLength, EntropyPower, and LengthPower(β).

As the results show, VN-Okapi gives improvements; however, the magnitude of these improvements is smaller than that in the case of VN-DP. One possible reason for the smaller improvement is the approximate two-stage normalization carried out in Okapi, as discussed in Section 2. As such, a form of verbosity normalization is performed by Okapi, to some extent, using the component $tfn/(tfn + 1)$, which causes VN-Okapi to have only a limited effect on retrieval performance.

Unlike in the case of DP, there is no significant difference between improvements for short keyword queries and verbose queries. Therefore, the argument made for DP wherein H2 is important, particularly for verbose queries, is much weaker for Okapi. Again, this is because Okapi has its own component $tfn/(tfn+1)$ that performs a form of normalization of verbose documents. As such, excessive preference for a verbose



(a) ROBUST (MAP) (b) ROBUST (P@5)

(c) WT10G (MAP) (d) WT10G (P@5)

(e) GOV2 (MAP) (f) GOV2 (P@5)

Fig. 1. Performance curves (MAP and P@5) of DP and VN-DP obtained using EntropyPower with varying μ in ROBUST (top), WT10G (center), and GOV2 (bottom).

document is handled to some extent by the original model, even without our explicit verbosity normalization.

The comparison results for three scope measures are also somewhat different from those of DP. In VN-Okapi, there is no winning scope measure between EntropyPower and UniqLength; in most cases, both have similar performance.

Table XI. The comparison of the best performance results for Okapi and VN-Okapi using EntropyPower, and the corresponding parameter values (b and k_1). The symbols * indicate that a run of the VN-Okapi shows statistically significant improvement over the baseline in the t-test at 0.95 confidence level.

		ROBUST	WT10G	GOV2
sk	Okapi	0.2454	0.2033	0.2920
		(0.3, 0.6)	(0.3, 0.6)	(0.01, 0.5)
	VN-Okapi	**0.2482***	**0.2107**	**0.3018***
		(0.1, 0.5)	(0.05, 0.3)	(0.1, 0.3)
sv	Okapi	0.2267	0.1935	0.2515
		(0.5, 1.0)	(0.5, 2.0)	(0.02, 0.6)
	VN-Okapi	**0.2303***	**0.2001**	**0.2618***
		(0.3, 0.6)	(0.3, 1.5)	(0.2, 0.6)
lv	Okapi	0.2637	0.2385	0.3012
		(0.8, 0.8)	(0.5, 1.5)	(0.03, 0.5)
	VN-Okapi	**0.2676***	**0.2482**	**0.3074***
		(0.4, 0.6)	(0.3, 0.8)	(0.3, 0.5)

Despite the limited effects, the improvements obtained by VN-Okapi are statistically significant, at least for either UniqLength or EntropyPower, for most of the collections and query types, and thus, they indicate the merit of our two-stage normalization.

For further comparison, Table XI presents the best MAPs for Okapi and VN-Okapi and their corresponding parameter values of b and k_1 for all three test collections and three query types.

A comparison of the optimal ranges of b across collections for both methods indicates that VN-Okapi tends to be robust without significant differences across collections, whereas Okapi has poor robustness with the optimal values of b being different between GOV2 and other collections. More specifically, for Okapi, the performance surfaces on GOV2 are shifted in the decreasing direction of b, relative to those of other collections. As a result, the best performance values of b become much smaller on GOV2 than on other collections for each query type; for short keyword queries, the best value of b is 0.01 on GOV2, and this is smaller than the value of 0.3 on other collections; a similar difference is observed for verbose queries. In contrast, for VN-Okapi, the parameter sensitivity of b on GOV2 is highly similar to that of other collections. The best performance values of b are not different across all collections; the best values of b are commonly between 0.05 and 0.1, for short keyword queries, between 0.2 and 0.3 for short verbose queries, and between 0.3 and 0.4 for long verbose queries.

A comparison of the best performances indicates that VN-Okapi is slightly better than Okapi, in that it highlights the small magnitude of the increase in MAP. Despite its small magnitude, on ROBUST and GOV2, the improvements over Okapi using VN-Okapi are statistically significant for all three types of queries.

6.3. MRF vs. VN-MRF

For evaluating MRF and VN-MRF, because we adopt sequential dependence, a dependency link (undirected link) is inserted only between two *adjacent* query words. Unlike in the case of other query types, for a long verbose query, we do not put a dependency across different topic fields. Thus, no dependency appears between a query word in the title field and a query word in the description or the narrative fields.

Table XII shows the comparative results of MRF and VN-MRF under three different scope measures, relative to those of DP and VN-DP using EntropyPower.

It is clearly seen that MRF is always better than DP, with all of the performance improvements being statistically significant. This precisely reproduces the comparison results reported by the existing works on MRF [Metzler and Croft 2005]. Note that

Table XII. MAP performance comparison of MRF and VN-MRF on three collections, ROBUST, WT10G, and GOV2, relative to that of DP and VN-DP. The symbols α, β, and γ indicate that a run of the VN method shows statistically significant improvement in the t-test at 0.95 confidence level, over DP, VN-DP, and MRF, respectively.

	Method	MRF (or VN-MRF)		
		ROBUST	WT10G	GOV2
sk	baseline (DP)	0.2447	0.1963	0.2907
	baseline (VN-DP)	0.2481^{α}	0.2120^{α}	0.3099^{α}
	baseline (MRF)	$0.2545^{\alpha\gamma}$	0.2149^{α}	0.3095^{α}
	LengthPower(0.5)	0.2506	0.2055	0.3032
	LengthPower(0.75)	$0.2557^{\alpha\gamma}$	0.2128^{α}	$0.3133^{\alpha\gamma}$
	LengthPower(0.9)	$0.2545^{\alpha\gamma}$	0.2142^{α}	$0.3125^{\alpha\gamma}$
	UniqLength	$0.2572^{\alpha\beta\gamma}$	$0.2244^{\alpha\gamma}$	$0.3270^{\alpha\beta\gamma}$
	EntropyPower	$\mathbf{0.2581}^{\alpha\beta\gamma}$	$\mathbf{0.2296}^{\alpha\beta\gamma}$	$\mathbf{0.3334}^{\alpha\beta\gamma}$
sv	baseline (DP)	0.2260	0.1909	0.2455
	baseline (VN-DP)	0.2440^{α}	0.2196^{α}	$0.2826^{\alpha\gamma}$
	baseline (MRF)	0.2416^{α}	0.2063^{α}	0.2687^{α}
	LengthPower(0.5)	$0.2545^{\alpha\beta\gamma}$	0.2197^{α}	$0.2810^{\alpha\gamma}$
	LengthPower(0.75)	$0.2507^{\alpha\beta\gamma}$	$0.2147^{\alpha\gamma}$	$0.2782^{\alpha\gamma}$
	LengthPower(0.9)	$0.2458^{\alpha\beta\gamma}$	$0.2125^{\alpha\gamma}$	$0.2739^{\alpha\gamma}$
	UniqLength	$0.2500^{\alpha\beta\gamma}$	$0.2214^{\alpha\gamma}$	$0.2879^{\alpha\gamma}$
	EntropyPower	$\mathbf{0.2550}^{\alpha\beta\gamma}$	$\mathbf{0.2368}^{\alpha\beta\gamma}$	$\mathbf{0.2975}^{\alpha\beta\gamma}$
lv	baseline (DP)	0.2707	0.2469	0.2864
	baseline (VN-DP)	0.2799^{α}	0.2614^{α}	0.3248^{α}
	baseline (MRF)	0.2813^{α}	0.2613^{α}	0.3164^{α}
	LengthPower(0.5)	$0.2866^{\alpha\beta\gamma}$	0.2581	$0.3368^{\alpha\beta\gamma}$
	LengthPower(0.75)	$0.2883^{\alpha\beta\gamma}$	0.2659^{α}	$0.3280^{\alpha\gamma}$
	LengthPower(0.9)	$0.2861^{\alpha\beta\gamma}$	0.2617^{α}	$0.3214^{\alpha\gamma}$
	UniqLength	$0.2895^{\alpha\beta\gamma}$	$0.2687^{\alpha\gamma}$	$0.3363^{\alpha\beta\gamma}$
	EntropyPower	$\mathbf{0.2927}^{\alpha\beta\gamma}$	$\mathbf{0.2754}^{\alpha\beta\gamma}$	$\mathbf{0.3481}^{\alpha\beta\gamma}$

the improved performance using MRF is further enhanced by VN-MRF with the application of the two-stage normalization, and additional improvements are statistically significant improvements. In particular, either on UniqLength or EntropyPower, VN-MRF is always better than MRF for all test collections and all query types, with all improvements being statistically significant.

Interestingly, VN-DP alone without exploiting the term dependency is nearly comparable to MRF, often even showing better performances. Again, the performance of VN-DP is further increased by VN-MRF along with the utilization of the term dependency, and the additional improvements are statistically significant in most cases, at least using either EntropyPower or UniqLength. Therefore, this result strongly implies that both effects resulting from the term dependency and the two-stage normalization are slightly co-related, thus facilitating such incremental increase by their combined utilization.

Another interesting result is that the performance difference of VN-MRF across test collections and query types shows a highly similar tendency to that of VN-DP. First, both VN methods (VN-MRF and VN-DP) are more effective, especially on the heterogeneous web collections (WT10G and GOV2) than on ROBUST. Second, on EntropyPower, both VN methods show larger improvements for verbose queries than for keyword queries – the only exception is found in VN-MRF for long verbose query on WT10G, where the improvement is slightly smaller than that for short keyword queries. Third, on LengthPower, both VN methods often show improvements over their original methods, and they are more effective for verbose queries than for keyword queries.

Table XIII. Comparison of performance of P@5 of MRF and VN-MRF for three collections, ROBUST, WT10G, and GOV2, relative to that of DP and VN-DP. The symbols α, β, and γ indicate that a run of the VN method shows statistically significant improvement in the t-test at 0.95 confidence level, over DP, VN-DP, and MRF, respectively.

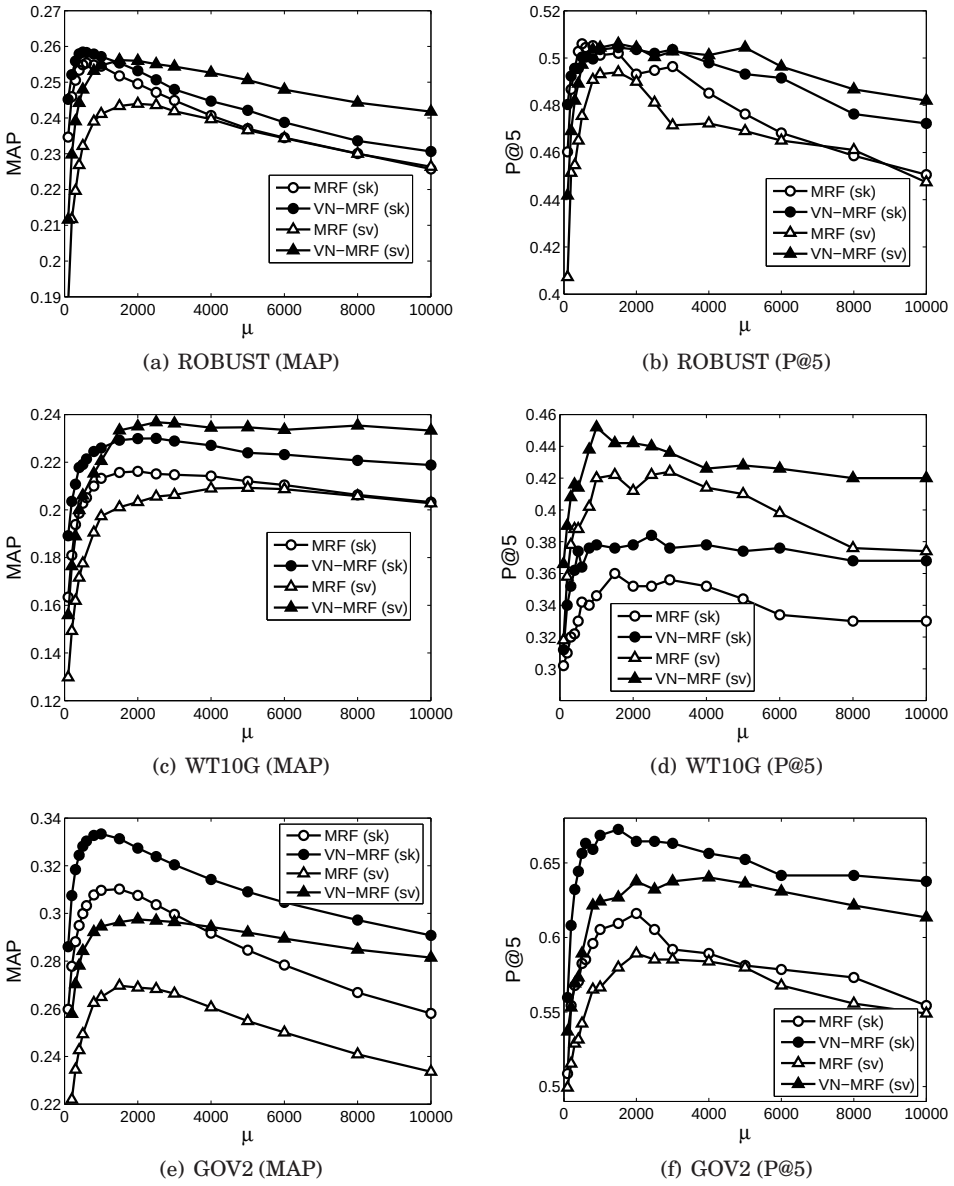
	Method	MRF (or VN-MRF)		
		ROBUST	WT10G	GOV2
sk	baseline (DP)	0.4924	0.3120	0.5678
	baseline (VN-DP)	0.4972	0.3640^{α}	0.6416
	baseline (MRF)	0.5036	0.3580^{α}	0.6121
	LengthPower(0.5)	0.4859	0.3540^{α}	0.5664
	LengthPower(0.75)	0.4916	0.3500^{α}	0.6054^{α}
	LengthPower(0.9)	0.5004	0.3540^{α}	0.6121^{α}
	UniqLength	**0.5068**	0.3660^{α}	$0.6470^{\alpha\gamma}$
	EntropyPower	0.5012	$\mathbf{0.3840}^{\alpha\beta\gamma}$	$\mathbf{0.6685}^{\alpha\gamma}$
sv	baseline (DP)	0.4466	0.3880	0.5208
	baseline (VN-DP)	0.4932^{α}	0.4300^{α}	0.6309
	baseline (MRF)	0.4876^{α}	0.4240^{α}	0.5839
	LengthPower(0.5)	0.4916^{α}	0.4140	0.5544
	LengthPower(0.75)	0.4972^{α}	0.4160	0.5785^{α}
	LengthPower(0.9)	0.4892^{α}	0.4120	0.5812^{α}
	UniqLength	0.4940^{α}	0.4320^{α}	$0.5919^{\alpha\gamma}$
	EntropyPower	$\mathbf{0.5044}^{\alpha\gamma}$	$\mathbf{0.4400}^{\alpha}$	$\mathbf{0.6376}^{\alpha\gamma}$
lv	baseline (DP)	0.5414	0.4460	0.6228
	baseline (VN-DP)	0.5631^{α}	0.4700	0.6644^{α}
	baseline (MRF)	0.5598^{α}	0.4700^{α}	0.6550^{α}
	LengthPower(0.5)	0.5582	0.4680	0.6336
	LengthPower(0.75)	0.5639^{α}	0.4720^{α}	0.6376
	LengthPower(0.9)	0.5655^{α}	0.4580	0.6456
	UniqLength	$0.5751^{\alpha\gamma}$	$\mathbf{0.4760}^{\alpha}$	0.6671^{α}
	EntropyPower	$\mathbf{0.5799}^{\alpha\beta\gamma}$	0.4640	$\mathbf{0.6886}^{\alpha\gamma}$

The similarity between the two VN methods is understandable considering the fact that the underlying retrieval function in MRF is basically the same as that of DP – DP and MRF commonly employ the smoothed document model of Eq. (6) for scoring a document.

Table XIII shows the performances of P@5 of VN-MRF, in comparison to those of MRF, using the values of the MAP-optimized parameters. The results for P@5 are largely similar to those for MAP, as seen in Table XII. In many cases, the MRF's performance of P@5 is better than that of DP, often with statistically significant improvements. Further, the performance of MRF is increased by VN-MRF with two-stage normalization, at least using either UniqLength or EntropyPower, and often with statistically significant improvements. In the particular case using EntropyPower, VN-MRF yields statistically significant improvements over MRF on WT10G and GOV2 for all short keyword queries and on ROBUST and GOV2 for some verbose queries. This result implies that in many cases, VN-MRF's significant improvement in MAP results from the increased performance of P@5. VN-DP alone shows performance similar to that of MRF. Again, the precision of VN-DP is slightly increased by VN-MRF exploiting the term dependency, although usually not to a statistically significant degree, unlike the results in MAP.

For further comparison, Figure 2 shows the performance curves of MRF and VN-MRF, plotted by varying μ, for short keyword and verbose queries – EntropyPower is used as the scope measure, and MAP and P@5 are used as the evaluation measures. Again, there is a great degree of similarity between the comparison results of VN-MRF and MRF and those of VN-DP and DP – for P@5 curves, VN-MRF is always better than the original method, except for only a few parameter values of μs in ROBUST. The

shapes of the performance curves of P@5 are quite similar for both VN-MRF and MRF. The optimal ranges of μ are also close, as in the case of VN-DP and DP.



(a) ROBUST (MAP) (b) ROBUST (P@5)

(c) WT10G (MAP) (d) WT10G (P@5)

(e) GOV2 (MAP) (f) GOV2 (P@5)

Fig. 2. Performance curves (MAP and P@5) of MRF and VN-MRF obtained using EntropyPower with varying μ on ROBUST (top), WT10G (center), and GOV2 (bottom).

7. APPLICATION TO LOWER BOUNDING TERM FREQUENCY NORMALIZATION

7.1. Lower-Bounded Retrieval Models

As discussed in related works, [Lv and Zhai 2009a] recently proposed the use of lower-bounding term frequency normalization in avoiding over-penalization of very long documents. Their experimental results showed that lower-bounded retrieval models lead to significant improvements in comparison with baseline models. An interesting issue is whether our proposed two-stage normalization can further improve these lower-bounded models. To this end, we chose DP and VN-DP as retrieval models, and compared their lower-bounded models with their VN models.

Two lower-bounded models for DP and VN-DP are presented in the following. First, a lower-bounded model for DP can be formulated as follows:

$$\sum_{w\in q\cap d} c(w,q)\left[\ln\left(1+\frac{c(w,d)}{\mu P(w|C)}\right)+\ln\left(1+\frac{\delta}{\mu P(w|C)}\right)\right]+|q|\ln\left(\frac{\mu}{|d|+\mu}\right) \tag{16}$$

which is called **DP+**. In Eq. (16), δ is a pseudo term frequency value that controls the scale of the lower bound, which was introduced by [Lv and Zhai 2011b].

Second, a lower-bounded model for VN-DP can be formulated by straightforwardly applying the general normalization approach of [Lv and Zhai 2011b][14].

$$\sum_{w\in q\cap d} c(w,q)\left[\ln\left(1+\frac{c(w,d)}{\mu P(w|C)}\frac{s(d)}{|d|}\right)+\ln\left(1+\frac{\delta}{\mu P(w|C)}\right)\right] +|q|\ln\left(\frac{\mu}{s(d)+\mu}\right) \tag{17}$$

which is called **VN-DP+**.

Similarly, we can derive a lower-bounded Okapi (**Okapi+**) and a lower-bounded VN-Okapi (**VN-Okapi+**). In the original BM25 retrieval formula, Okapi+ uses $tf_{BM25+}(w,d)$ (i.e., $tf_{BM25}(w,d)$ + δ) for the term frequency component, and VN-Okapi+ uses $tf_{BM25}(w,\phi(d))$ (i.e., $tf_{BM25}(w,\phi(d))$ + δ) which are given by

$$tf_{BM25+}(w,d)=\frac{(k_1+1)c(w,d)}{k_1\left((1-b)+b|d|/avgl\right)+c(w,d)}+\delta \tag{18}$$

$$tf_{BM25+}(w,\phi(d))=\frac{(k_1+1)c(w,d)}{k_1|d|\left(((1-b)/s(d))+b/avgs\right)+c(w,d)}+\delta \tag{19}$$

7.2. Experiment Results

For parameter tuning of the lower-bounded models, we follow the setting in the work of [Lv and Zhai 2009a]; For DP+ and VN-DP+, we search δ over the space between 0 and 0.15, with increments of 0.01. For Okapi+ and VN-Okapi+, we search δ over space between 0.0 and 1.5, with the increment of 0.1. The search spaces for other retrieval parameters are the same as those used in previous sections.

Table XIV shows the MAP performances of DP+ and VN-DP+, as compared to DP and VN-DP. As shown in Table XIV, DP+ often exhibits non-trivial improvements over DP, especially for short verbose queries, reaffirming the results reported in [Lv and Zhai 2011b] that DP+ shows greater effectiveness for short verbose queries than for keyword queries, which shows a different result from that achieved by

[14]According to the notation of [Lv and Zhai 2011b], $F(c(w,\phi(d)),|\phi(d)|,td(w))$ corresponds to $\ln\left(\frac{\mu}{s(d)+\mu}+\frac{c(w,\phi(d))}{(s(d)+\mu)P(w|C)}\right)$

Table XIV. MAP performance comparison of DP and VN-DP on three collections ROBUST, WT10G, and GOV2, and three different query types sk, sv, and lv. EntropyPower is used for the scope measure in VN-DP and VN-DP+. Symbols α, β, and γ indicate that a run of the VN method (or a lower-bounded method) shows a statistically significant improvement over DP, DP+, VN-DP, respectively, in the t-test at 0.95 confidence level.

	Method	DP+ (or VN-DP+)		
		ROBUST	WT10G	GOV2
sk	DP	0.2447	0.1963	0.2907
	DP+	0.2447	0.1957	0.2922
	VN-DP	**0.2481**$^{\alpha\beta}$	**0.2120**$^{\alpha\beta}$	0.3099$^{\alpha\beta}$
	VN-DP+	0.2476$^{\alpha\beta}$	0.2112$^{\alpha\beta}$	**0.3141**$^{\alpha\beta\gamma}$
sv	DP	0.2260	0.1909	0.2455
	DP+	0.2337$^{\alpha}$	0.1969	0.2453
	VN-DP	0.2440$^{\alpha\beta}$	0.2196$^{\alpha\beta}$	**0.2826**$^{\alpha\beta}$
	VN-DP+	**0.2461**$^{\alpha\beta}$	**0.2215**$^{\alpha\beta}$	0.2819$^{\alpha\beta}$
lv	DP	0.2707	0.2469	0.2864
	DP+	0.2766	0.2442	0.2863
	VN-DP	0.2799$^{\alpha}$	**0.2614**$^{\alpha\beta}$	0.3248$^{\alpha\beta}$
	VN-DP+	**0.2858**$^{\alpha\beta\gamma}$	0.2603$^{\alpha\beta}$	**0.3248**$^{\alpha\beta}$

[Lv and Zhai 2011b]; our experiment shows a statistically significant improvement when using DP+ over DP for only sv queries in the ROBUST collection. Unlike DP+, VN-DP+, the lower-bounded model over VN-DP, does not show greater effectiveness for short verbose queries than for other types of queries. This may be because VN-DP already shows a significant improvement over DP for short verbose queries, and further improvement is therefore less likely. Nevertheless, VN-DP+ continues to further increase the performances of VN-DP, with improvements being statistically significant for short keyword queries in GOV2 and long verbose queries in ROBUST.

Importantly, on comparing VN-DP with DP+, we can see that DP+ does not reach the performance of VN-DP. For almost all runs (except for lv in ROBUST), the improvements gained by VN-DP over DP are mostly larger than those made by DP+ over DP, and in most cases are statistically significant. Furthermore, VN-DP leads mostly to statistically significant improvements over DP+ for almost all runs. These results clearly demonstrate that the improvement from the VN model over DP is not redundant to the effects from the existing lower-bounding normalization, and leads to a significant improvement even against lower-bounded models, which are stronger baselines. Overall, our experimental results indicate that two-stage normalization significantly improves lower-bounded models for almost all runs for three different collections.

We now consider the comparison between lower-bounded models for Okapi and VN-Okapi and their original models. Table XV lists the MAP performances of Okapi+ and VN-Okapi+, as compared to those of Okapi and VN-Okapi. Again, results similar to those presented in Table XIV are obtained, although the improvements by the VN models over lower-bounded models are not larger than the case of DP; the lower-bounded models are effective in improving baseline models, without reaching the performance of VN-Okapi. The improvements gained by VN-Okapi over Okapi are mostly larger than those made by Okapi+ over Okapi. Although the improvements of VN-Okapi over Okapi+ are not statistically significant in most cases, VN-Okapi+ leads to statistically significant improvements over Okapi+ for almost all runs.

For further comparison, we present the performances of original, VN, and lower-bounded models with respect to standard topic sets of TREC in three test collections, named TREC6, TREC7, TREC8, ROBUST03, ROBUST04, TREC9, TREC10,

Table XV. MAP performance comparison of Okapi and VN-Okapi on three collections ROBUST, WT10G, and GOV2, and three different query types sk, sv, and lv. EntropyPower is used for the scope measure in VN-Okapi and VN-Okapi+. Symbols α, β, and γ indicate that a run of the VN method (or a lower-bounded method) shows a statistically significant improvement over Okapi, Okapi+, VN-Okapi, respectively, in the t-test at 0.95 confidence level.

	Method	Okapi+ (or VN-Okapi+)		
		ROBUST	WT10G	GOV2
sk	Okapi	0.2444	0.1946	0.2920
	Okapi+	0.2457^{α}	0.2039^{α}	0.2969^{α}
	VN-Okapi	0.2477^{α}	0.2071^{α}	0.3004^{α}
	VN-Okapi+	$\mathbf{0.2477}^{\alpha}$	$\mathbf{0.2085}^{\alpha}$	$\mathbf{0.3100}^{\alpha\beta\gamma}$
sv	Okapi	0.2247	0.1884	0.2498
	Okapi+	0.2279^{α}	0.1900	0.2573^{α}
	VN-Okapi	0.2303^{α}	0.1968^{α}	0.2599^{α}
	VN-Okapi+	$\mathbf{0.2311}^{\alpha\beta}$	$\mathbf{0.2023}^{\alpha\gamma}$	$\mathbf{0.2658}^{\alpha\beta\gamma}$
lv	Okapi	0.2619	0.2314	0.3012
	Okapi+	0.2640^{α}	0.2320	0.3059^{α}
	VN-Okapi	$\mathbf{0.2659}^{\alpha}$	$\mathbf{0.2415}^{\alpha\beta}$	0.3074^{α}
	VN-Okapi+	0.2658^{α}	$0.2390^{\alpha\beta}$	$\mathbf{0.3094}^{\alpha\beta}$

Table XVI. Standard topic sets of TREC, their corresponding collection names, and their training topic sets.

Topic set id	Query ids	Collection	Training topic sets
TREC6	Q301-Q350	ROBUST	TREC7,TREC8,ROBUST03,ROBUST04
TREC7	Q351-Q400		TREC6,TREC8,ROBUST03,ROBUST04
TREC8	Q401-Q450		TREC6,TREC7,ROBUST03,ROBUST04
ROBUST03	Q601-Q650		TREC6,TREC7,TREC8,ROBUST04
ROBUST04	Q651-Q700		TREC6,TREC7,TREC8,ROBUST03
TREC9	Q451-Q500	WT10G	TREC10
TREC10	Q501-Q550		TREC9
TREC2004	Q701-Q750	GOV2	TREC2005,TREC2006
TREC2005	Q751-Q800		TREC2004,TREC2006
TREC2006	Q801-Q850		TREC2004,TREC2005

TREC2004, TREC2005, and TREC2006. Table XVI presents the basic information on the standard topic sets of TREC.

First, Table XVII shows the MAP performances between DP+ and VN-DP+, as compared to DP and VN-DP on standard TREC topic sets. As shown in Table XVII, VN-DP or VN-DP+ show further improvements over DP and DP+ for almost all standard topic sets and they are statistically significant for more than half of all cases. In particular, VN-DP+ shows the best performance for almost all runs. Their improvements over DP are statistically significant (except for standard topic sets of sk queries in ROBUST and lv queries in WT10G) and are larger than the improvements of VN-DP or DP+ over DP. Comparing VN-DP to DP+, more runs showed improvements of statistical significance on VN-DP over DP than on DP+ over DP.

Turning to the comparison between lower-bounded and VN models for Okapi, Table XVIII shows the MAP performances between Okapi+ and VN-Okapi+, as compared to Okapi and VN-Okapi on standard TREC topic sets. Again, VN-Okapi+ shows the best performance for almost all runs, and their improvements over Okapi are larger than those made by VN-Okapi or Okapi+ over Okapi, being statistically significant for most cases.

Thus, the main results of Tables XVII and XVIII are largely consistent with those reported in Tables XIV and XV, respectively. The lower bounding models do not reach the performance of VN models; the improvements gained by VN models over the baseline are mostly larger than those made by lower bounding models over the baseline; the fur-

Table XVII. MAP performance comparison of DP, DP+, VN-DP, and VN-DP+ on standard topic sets in TREC and three different query types sk, sv, and lv. EntropyPower is used for the scope measure in VN-DP and VN-DP+. Symbols α, β, and γ indicate that a run of the VN method (or a lower-bounded method) shows a statistically significant improvement over DP, DP+, VN-DP, respectively, in the t-test at 0.95 confidence level.

		DP	DP+	VN-DP	VN-DP+
sk	TREC6	0.2465	0.2471	**0.2483**	0.2479
	TREC7	0.1733	0.1733	**0.1785**	0.1783
	TREC8	0.2410	0.2415	0.2425	**0.2425**
	ROBUST03	0.2756	0.2755	**0.2818**$^{\alpha\beta}$	0.2812$^{\alpha\beta}$
	ROBUST04	0.2879	0.2871	**0.2903**	0.2892
	TREC9	0.1985	0.1997	0.2063	**0.2068**
	TREC10	0.1942	0.1917	**0.2177**$^{\alpha\beta}$	0.2156$^{\alpha\beta}$
	TREC2004	0.2597	0.2605	0.2790$^{\alpha\beta}$	**0.2842**$^{\alpha\beta}$
	TREC2005	0.3114	0.3130	0.3297$^{\alpha\beta}$	**0.3336**$^{\alpha\beta}$
	TREC2006	0.3005	0.3026	0.3205$^{\alpha\beta}$	**0.3228**$^{\alpha\beta}$
sv	TREC6	0.1751	0.1898	0.1980$^{\alpha}$	**0.2018**$^{\alpha\beta}$
	TREC7	0.1698	0.1768$^{\alpha}$	0.1887$^{\alpha\beta}$	**0.1904**$^{\alpha\beta}$
	TREC8	0.2145	0.2194	0.2301	**0.2306**$^{\alpha\beta}$
	ROBUST03	0.2912	0.2996	0.3171$^{\alpha\beta}$	**0.3182**$^{\alpha\beta}$
	ROBUST04	0.2806	0.2840	0.2872	**0.2906**$^{\alpha}$
	TREC9	0.1996	0.2094$^{\alpha}$	0.2299	**0.2325**$^{\alpha}$
	TREC10	0.1822	0.1844	0.2093$^{\alpha\beta}$	**0.2105**$^{\alpha\beta}$
	TREC2004	0.2163	0.2154	**0.2498**$^{\alpha\beta}$	0.2475$^{\alpha\beta}$
	TREC2005	0.2524	0.2524	0.2860$^{\alpha\beta}$	**0.2860**$^{\alpha\beta}$
	TREC2006	0.2671	0.2676	0.3114 $^{\alpha\beta}$	**0.3114**$^{\alpha\beta}$
lv	TREC6	0.2627	0.2792$^{\alpha}$	0.2713$^{\alpha}$	**0.2904**$^{\alpha\beta}$
	TREC7	0.2152	0.2254$^{\alpha}$	0.2219$^{\alpha}$	**0.2296**$^{\alpha\beta}$
	TREC8	0.2421	0.2511$^{\alpha}$	0.2539$^{\alpha}$	**0.2619**$^{\alpha\beta\gamma}$
	ROBUST03	0.3289	0.3252	0.3391$^{\alpha\beta}$	**0.3401**$^{\alpha\beta}$
	ROBUST04	0.3051	0.3063	**0.3106**	0.3074
	TREC9	0.2550	0.2550	**0.2675**	0.2675
	TREC10	0.2388	0.2334	**0.2553**$^{\alpha\beta}$	0.2530
	TREC2004	0.2650	0.2651	0.2943$^{\alpha\beta}$	**0.2943**$^{\alpha\beta}$
	TREC2005	0.2875	0.2870	0.3232$^{\alpha\beta}$	**0.3232**$^{\alpha\beta}$
	TREC2006	0.3062	0.3062	0.3563$^{\alpha\beta}$	**0.3563**$^{\alpha\beta}$

ther improvements even against lower bounding models are made by lower bounding VN models (VN-DP+ or VN-Okapi+) [15].

As a consequence, the overall results shown in Table XIV, XV, XVII, and XVIII consistently indicate that the improvement resulting from the application of two-stage normalization is not fully replaceable by adopting lower-bounding term frequency normalization, and vice versa. This result is intuitive, as two normalizations aim at different deficiencies of the existing normalization method: lower-bounding term frequency normalization aims at avoiding the penalization of *very* long documents, while two-

[15] However, compared to the previous experiments, the statistical significance for VN models is weakly supported on some sets of standard queries for both the DP and Okapi cases. The result of the statistical significance is also fairly observed in lower bounding models where the improvements are not statistically significant on some of the standard TREC topic sets. We believe that the reason is the lack of evidence for by which to judge significance. The number of queries used in standard topics is 50, which is often not sufficient to convincingly decide significance, especially when the improvement is marginal. Consider, for example, the case of Okapi and sk queries in ROBUST (shown in Table XVIII). When using only 50 queries in standard TREC, the improvements made by VN-Okapi and Okapi+ over the baseline are mostly not of statistical significance. However, when using 250 queries in ROBUST, the improvements turn out to be statistically significant as shown in Table XV. Thus, the results show that a larger number of queries might be necessary for the statistical significance test when the improvements are marginal.

Table XVIII. MAP performance comparison of Okapi, Okapi+, VN-Okapi, and VN-Okapi+ on standard topic sets in TREC and three different query types sk, sv, and lv. EntropyPower is used for the scope measure in VN-Okapi and VN-Okapi+. Symbols α, β, and γ indicate that a run of the VN method (or a lower-bounded method) shows a statistically significant improvement over Okapi, Okapi+, VN-Okapi, respectively, in the t-test at 0.95 confidence level.

		Okapi	Okapi+	VN-Okapi	VN-Okapi+
sk	TREC6	0.2471	0.2479	0.2492	**0.2498**
	TREC7	0.1734	0.1754	0.1758	**0.1773**$^{\alpha}$
	TREC8	0.2413	0.2415	**0.2445**	0.2416
	ROBUST03	0.2760	0.2786	0.2826$^{\alpha}$	**0.2833**$^{\alpha}$
	ROBUST04	0.2848	0.2861	0.2872	**0.2874**
	TREC9	0.1946	0.2077$^{\alpha}$	0.2050	**0.2089**
	TREC10	0.1946	0.2001	**0.2092**$^{\alpha}$	0.2081
	TREC2004	0.2562	0.2606	0.2659	**0.2766**$^{\alpha\beta\gamma}$
	TREC2005	0.3222	0.3297$^{\alpha}$	0.3233	**0.3354**$^{\alpha\gamma}$
	TREC2006	0.2969	0.2996	0.3113$^{\alpha}$	**0.3174**$^{\alpha\beta}$
sk	TREC6	0.1628	0.1663	0.1678	**0.1678**
	TREC7	0.1603	0.1642$^{\alpha}$	0.1653$^{\alpha}$	**0.1702**$^{\alpha\beta\gamma}$
	TREC8	0.2111	0.2144	0.2154	**0.2172**$^{\alpha}$
	ROBUST03	0.3148	0.3145	**0.3196**$^{\alpha}$	0.3159
	ROBUST04	0.2753	0.2809$^{\alpha}$	0.2847	**0.2847**$^{\alpha}$
	TREC9	0.1958	0.1970	0.2055	**0.2117**$^{\alpha\gamma}$
	TREC10	0.1809	0.1831	0.1880	**0.1930**$^{\alpha}$
	TREC2004	0.2240	0.2302	0.2327$^{\alpha}$	**0.2401**$^{\alpha\beta\gamma}$
	TREC2005	0.2540	0.2682$^{\alpha}$	0.2597	**0.2714**$^{\alpha\gamma}$
	TREC2006	0.2707	0.2731	**0.2866**$^{\alpha}$	0.2855$^{\alpha\beta}$
lv	TREC6	0.2365	0.2360	0.2366	**0.2374**
	TREC7	0.2077	**0.2104**$^{\alpha}$	0.2093	0.2093
	TREC8	0.2420	0.2434	0.2448	**0.2451**
	ROBUST03	0.3287	0.3331	0.3378	**0.3381**$^{\alpha}$
	ROBUST04	0.2954	0.2980	**0.3016**$^{\alpha}$	0.3000$^{\alpha}$
	TREC9	0.2248	0.2281$^{\alpha}$	**0.2413**$^{\alpha\beta}$	0.2362$^{\alpha\beta}$
	TREC10	0.2379	0.2359	0.2418	**0.2418**
	TREC2004	0.2695	0.2710	0.2743	**0.2746**
	TREC2005	0.3056	0.3146$^{\alpha}$	0.3083	**0.3164**$^{\alpha}$
	TREC2006	0.3280	0.3315	**0.3389**$^{\alpha}$	0.3365$^{\alpha\beta}$

stage normalization aims at avoiding *insufficient* penalization of verbose documents and *excessive* penalization of long documents. The resultant solutions for these different goals also differ: lower-bounding term frequency normalization does not need to decompose the document length into additional factors, but rather enforces the addition of scope gaps between document scores when a query term appears and disappears in a document. Two-stage normalization decomposes a document length into verbosity and scope factors, penalizing verbose and broad documents *separately*.

In conclusion, given the set of results described throughout this section, we can state the following. The proposed two-stage normalization is clearly effective for further improving the existing retrieval model; it is not limitedly applicable to improving the baseline retrieval model and can also be extended even to the improved methods that uses the term dependency and the lower-bounding term frequency normalization. From the comparative axiomatic analysis results, we conclude that the normalization heuristics H1 and H2 should necessarily be applied for scoring a document.

8. CONCLUSION

In this paper, we argue that a normalization function should use different penalizations for verbosity and scope, and we propose the use of two-stage normalization. Our main contributions over and above those of previous works that formulated ranking

functions belonging to two-stage normalization [Singhal et al. 1996; Na et al. 2008a] are as follows: 1) We generalize two-stage normalization such that it can be applied to any retrieval model. 2) We perform comparative axiomatic analysis and capture the exact retrieval heuristics resulting from two-stage normalization and its difference from the original method. The results of experiments on three models – DP, Okapi, and MRF – consistently show that two-stage normalization is promising.

Of course, considerable work needs to be done in the future. Although two-stage normalization is effective in the case of DP, Okapi, and MRF, we still need more evidence of its effectiveness for other retrieval models. Thus, an obvious future work is to explore the application of two-stage normalization to the pivoted vector space model [Singhal et al. 1996], DFR [Amati and Van Rijsbergen 2002; Amati and Rijsbergen 2002], a more recently developed information model [Clinchant and Gaussier 2010], a parameterized query expansion [Bendersky et al. 2011], a *term-specific* adaptation of normalization parameter [Lv and Zhai 2011a], or a learning-to-rank framework [Liu 2009; Li 2011]. Another research direction is strengthening the axiomatic framework by generalizing the current retrieval constraints such that they can effectively cover the conjectured retrieval heuristics derived in this paper. A more challenging future research direction is to develop a new retrieval model in an innovative manner such that it includes the verbosity, scope, and document length as retrieval parameters.

REFERENCES

James Allan. 1995. Relevance feedback with too much data. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '95)*. 337–343.

James Allan, Margaret E. Connell, W. Bruce Croft, Fang-Fang Feng, David Fisher, and Xioayan Li. 2000. INQUERY and TREC-9. In *TREC-9*.

Gianni Amati and C. J. van Rijsbergen. 2002. Term Frequency Normalization via Pareto Distributions. In *Proceedings of the 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval (ECIR '02)*. 183–192.

Gianni Amati and Cornelis Joost Van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)* 20 (2002), 357–389. Issue 4.

Michael Bendersky and Oren Kurland. 2008. Utilizing passage-based language models for document retrieval. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval (ECIR'08)*. 162–174.

Michael Bendersky and Oren Kurland. 2010. Utilizing passage-based language models for ad hoc document retrieval. *Information Retrieval* 13 (April 2010), 157–187. Issue 2.

Michael Bendersky, Donald Metzler, and W. Bruce Croft. 2010. Learning concept importance using a weighted dependence model. In *Proceedings of the third ACM international conference on Web search and data mining (WSDM '10)*. 31–40.

Michael Bendersky, Donald Metzler, and W. Bruce Croft. 2011. Parameterized Concept Weighting in Verbose Queries. In *Proceedings of the 34th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '11)*.

James Callan. 1994. Passage-level evidence in document retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '94)*. 302–310.

James Callan, W. Bruce Croft, and Stephen M. Harding. 1992. The INQUERY Retrieval System. In *In Proceedings of the Third International Conference on Database and Expert Systems Applications*. 78–83.

Stéphane Clinchant and Éric Gaussier. 2010. Information-based models for ad hoc IR. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '10)*. 234–241.

Stéphane Clinchant and Éric Gaussier. 2011. A Document Frequency Constraint for Pseudo-Relevance Feedback Models. In *COnférence en Recherche d'Infomations et Applications - 8th French Information Retrieval Conference (CORIA '11)*. 73–88.

Stéphane Clinchant and Eric Gaussier. 2013. A Theoretical Analysis of Pseudo-Relevance Feedback Models. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval (ICTIR '13)*. 6:6–6:13.

Bruce Croft, Donald Metzler, and Trevor Strohman. 2009. *Search Engines: Information Retrieval in Practice* (1st ed.). Addison-Wesley Publishing Company.

Ronan Cummins and Colm O'Riordan. 2006. Evolved term-weighting schemes in Information Retrieval: an analysis of the solution space. *Artificial Intelligence Review* 26 (2006), 35–47. Issue 1-2.

Hui Fang, Tao Tao, and ChengXiang Zhai. 2004. A formal study of information retrieval heuristics. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04)*. 49–56.

Hui Fang, Tao Tao, and Chengxiang Zhai. 2011. Diagnostic Evaluation of Information Retrieval Models. *ACM Transactions on Information Systems (TOIS)* 29 (2011), 7:1–7:42. Issue 2.

Hui Fang and ChengXiang Zhai. 2005. An exploration of axiomatic approaches to information retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05)*. 480–487.

Hui Fang and ChengXiang Zhai. 2006. Semantic term matching in axiomatic approaches to information retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '06)*. 115–122.

Ben He and Iadh Ounis. 2003. A study of parameter tuning for term frequency normalization. In *Proceedings of the twelfth international conference on Information and knowledge management (CIKM '03)*. 10–16.

H. S. Heaps. 1978. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, Inc., Orlando, FL, USA.

Maryam Karimzadehgan and ChengXiang Zhai. 2012. Axiomatic Analysis of Translation Language Model for Information Retrieval. In *Proceedings of the 34th European Conference on Advances in Information Retrieval (ECIR'12)*. 268–280.

Marcin Kaszkiel and Justin Zobel. 1997. Passage retrieval revisited. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '97)*. 178–185.

Marcin Kaszkiel, Justin Zobel, and Ron Sacks-Davis. 1999. Efficient passage ranking for document databases. *ACM Transactions on Information Systems (TOIS)* 17 (1999), 406–439. Issue 4.

Eyal Krikon and Oren Kurland. 2011. A study of the integration of passage-, document-, and cluster-based information for re-ranking search results. *Information Retrieval* published online (May 2011).

Eyal Krikon, Oren Kurland, and Michael Bendersky. 2010. Utilizing inter-passage and inter-document similarities for reranking search results. *ACM Transactions on Information Systems (TOIS)* 29 (2010), 3:1–3:28. Issue 1.

Oren Kurland and Lillian Lee. 2005. PageRank without hyperlinks: structural re-ranking using links induced by language models. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05)*. 306–313.

Oren Kurland and Lillian Lee. 2009. Clusters, language models, and ad hoc information retrieval. *ACM Transactions on Information Systems (TOIS)* 27 (May 2009), 13:1–13:39. Issue 3.

Hao Lang, Donald Metzler, Bin Wang, and Jin-Tao Li. 2010. Improved latent concept expansion using hierarchical markov random fields. In *Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10)*. 249–258.

Matthew Lease. 2009. An improved markov random field model for supporting verbose queries. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '09)*. 476–483.

Hang Li. 2011. *Learning to Rank for Information Retrieval and Natural Language Processing*. Morgan & Claypool Publishers.

Tie-Yan Liu. 2009. Learning to Rank for Information Retrieval. *Foundations and Trends in Information Retrieval* 3 (March 2009), 225–331. Issue 3.

Xiaoyong Liu and W. Bruce Croft. 2002. Passage retrieval based on language models. In *Proceedings of the eleventh international conference on Information and knowledge management (CIKM '02)*. 375–382.

Yuanhua Lv and ChengXiang Zhai. 2009a. A comparative study of methods for estimating query language models with pseudo feedback. In *Proceeding of the 18th ACM conference on Information and knowledge management (CIKM '09)*. 1895–1898.

Yuanhua Lv and ChengXiang Zhai. 2009b. Positional language models for information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '09)*. 299–306.

Yuanhua Lv and ChengXiang Zhai. 2010. Positional relevance model for pseudo-relevance feedback. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '10)*. 579–586.

Yuanhua Lv and ChengXiang Zhai. 2011a. Adaptive term frequency normalization for BM25. In *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM '11)*. 1985–1988.

Yuanhua Lv and ChengXiang Zhai. 2011b. Lower-bounding term frequency normalization. In *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM '11)*. 7–16.

Yuanhua Lv and ChengXiang Zhai. 2011c. When documents are very long, BM25 fails!. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR '11)*. 1103–1104.

Donald Metzler and W. Bruce Croft. 2007. Linear feature-based models for information retrieval. *Information Retrieval* 10 (June 2007), 257–274. Issue 3.

Donald Metzler and W. Bruce Croft. 2005. A Markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05)*. 472–479.

Donald Metzler and W. Bruce Croft. 2007. Latent concept expansion using markov random fields. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '07)*. 311–318.

Elke Mittendorf and Peter Schäuble. 1994. Document and passage retrieval based on hidden Markov models. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '94)*. 318–327.

Seung-Hoon Na, In-Su Kang, and Jong-Hyeok Lee. 2008a. Improving term frequency normalization for multi-topical documents and application to language modeling approaches. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval (ECIR'08)*. 382–393.

Seung-Hoon Na, In-Su Kang, Ye-Ha Lee, and Jong-Hyeok Lee. 2008b. Completely-arbitrary passage retrieval in language modeling approach. In *Proceedings of the 4th Asia information retrieval conference on Information retrieval technology (AIRS'08)*. 22–33.

Stephen Robertson and Stephen Walker. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '94)*. 232–241.

Stephen Robertson, Stephen Walker, K. Sparck Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at TREC-3. In *Proceedings of the Thrid Text REtrieval Conference (TREC-3)*.

Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* 3, 4 (2009), 333–389.

Gerard Salton, J. Allan, and Chris Buckley. 1993. Approaches to passage retrieval in full text information systems. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '93)*. 49–58.

Gerard Salton and Chris Buckley. 1991. Automatic text structuring and retrieval-experiments in automatic encyclopedia searching. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '91)*. 21–30.

Amit Singhal, Chris Buckley, and Mandar Mitra. 1996. Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '96)*. 21–29.

Tao Tao and ChengXiang Zhai. 2007. An exploration of proximity measures in information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '07)*. 295–302.

Lidan Wang, Jimmy Lin, and Donald Metzler. 2010. Learning to efficiently rank. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '10)*. 138–145.

Chengxiang Zhai and John Lafferty. 2001. A study of smoothing methods for language models applied to Ad Hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01)*. 334–342.

Appendix A: Definition of Verbosity

For full definition of verbosity, suppose that M is the total number of *topics* in a collection, and $s(d)$ is the number of topics mentioned in d (or the *expected* number of topics in d). Here, we assume that the topic is *countable*, which may refer to an individual

word or a concept. Given document d, we first define the *topic-specific verbosity* of document d, noted $v(t,d)$, which is the sum of frequencies of all words which belong to t:

$$v(t,d) = \sum_{w \in \mathcal{V}} c(w,d)P(t|w) \tag{20}$$

where $P(t|w)$ is the posterior probability that w comes from t (i.e., $\sum_{i=1}^{M} P(t_i|w) = 1$).

Under Eq. (20), we readily show that $v(t,d)$ is the length of the passages in d which belong to t.

By the definition above, we further show that the following equality holds:

$$\begin{aligned} |d| &= \sum_{i=1}^{M} v(t_i,d) \\ &= v(t_1,d) + \cdots + v(t_M,d) \end{aligned} \tag{21}$$

To simplify Eq. (21), note that $v(t,d) = 0$ for most topics, as documents usually cover only a few topics. Let $i_1 \cdots i_k \cdots i_{s(d)}$ be indexes of topics appearing in d where $v(i_k,d) > 0$. Then, $|d|$ is reformulated as

$$|d| = v(t_{i_1},d) + \cdots + v(t_{i_{s(d)}},d) \tag{22}$$

Now, $v(d)$, the verbosity of document d, is defined as the average of all *per-topic* verbosities computed for all $s(d)$ topics appearing in d, which is given by:

$$\begin{aligned} v(d) &= \frac{v(t_{i_1},d) + \cdots + v(t_{i_{s(d)}},d)}{s(d)} \\ &= \frac{|d|}{s(d)} \end{aligned} \tag{23}$$

Thus, the average verbosity is the document length divided by the number of topics $s(d)$, which exactly replicates Eq. (2). Given Eq. (23), we only require $s(d)$, without need to estimate $v(t,d)$ which are usually unseen and hard to compute.

Appendix B: On Document-Specific Conjugate Prior for VN-DP

The use of a document-specific conjugate prior for VN-DP (i.e., Eq. (5)) is derived from the verbosity hypothesis. For convenience of discussion, suppose that $c'(w,d)$ indicates the unseen frequency of w in d. Generally, smoothing uses $c_s(w,d)$ defined as $c(w,d) + c'(w,d)$ as a count of w in d, thereby estimating $P(w|d)$ as $c_s(w,d)/\sum_{w \in V} c_s(w,d)$.

In DP, it is assumed that the *pseudo length* μ is distributed over unseen words according to $P(w|C)$, resulting in $c'(w,d) = \mu P(w|C)$. In our case, however, because document length is decomposed to verbosity and scope, we need to introduce two pseudo factors for unseen words: verbosity and scope. Formally, let $v'(d)$ and $s'(d)$ be the verbosity and scope of an unseen part of d, respectively. Just like the formula of frequencies of seen words, which is given by $c(w,d) = v(d)s(d)P_{ml}(w|d)$, frequencies of unseen words are formulated as $c'(w,d) = v'(d)s'(d)P(w|C)$.

To determine $v'(d)$ and $s'(d)$, we use the following assumptions:

1. *Verbosity of an unseen part: given a document, the verbosity of unseen passages (i.e., consisting of all unseen words) is the same as verbosity of the document.* – The assumption is due to the verbosity hypothesis; $c(w,d)$ is mostly governed by $v(d)$. Just as the verbosity hypothesis is applied to seen words, we apply the verbosity hypothesis to unseen words. This results in $c'(w,d)$, the frequencies of unseen words, which should also be governed by $v(d)$.

2. *Scope of an unseen part: given a document, the unseen scope of passages (i.e., consisting of all unseen words) is independent of the scope of the document.* – Unlike verbosity, we do not make a document-specific setting for the scope, as the relation between the unseen scope of d and $v(d)$ is not very clear.

Under these assumptions, we have $v'(d) = v(d)$, and $s'(d) = \mu$, thus resulting in $c'(w,d) = \mu v(d) P(w|C)$, which leads to our use of a document-specific prior in Eq. (5).

Therefore, the difference in formulating $c'(w,d)$ between DP and VN-DP results from whether or not we use the verbosity hypothesis for determining frequencies of unseen words.

Appendix C: Comparative Axiomatic Analysis

In this appendix, we briefly summarize the derivations of C_1, C_2, and C_3 for Okapi and DP, where C_1 and C_3 are necessary but not sufficient for satisfying the particular constraint. Let d_1 and d_2 be two given documents for LNCs and TF-LNC and $\Delta f(d_1, d_2, q)$ be $f(d_1, q) - f(d_2, q)$. All our derivations start from the inequality of $\Delta f(d_1, d_2, q) \geq 0$ (or $\Delta f(d1, d2, q) > 0$). For VN-DP, $\Delta f(d1, d2, q) \geq 0$ is equivalent to

$$\frac{s(d_1)p_{ml}(w|d_1) + \mu p(w|C)}{s(d_1) + \mu} \geq \frac{s(d_2)p_{ml}(w|d_1) + \mu p(w|C)}{s(d_2) + \mu} \tag{24}$$

For VN-Okapi, $\Delta f(d_1, d_2, q) \geq 0$ is equivalent to

$$\frac{c(w,d_1)}{v(d_1)} \left(k \left(1 - b + b \frac{s(d_1)}{avgs} \right) \right) idf(w) \geq \frac{c(w,d_2)}{v(d_2)} \left(k \left(1 - b + b \frac{s(d_2)}{avgs} \right) \right) idf(w) \tag{25}$$

1) LNC1

We first show the derivation of the conditions for LNC1 under VN-DP and VN-Okapi. For the sake of convenience, we introduce the variables m and ε that are defined as $m = v(d_1)/v(d_2)$ and $\varepsilon = K/|d_1|$, respectively. According to the definition of PAN, we have the following relation between $s(d_1)$ and $s(d_2)$:

$$\begin{aligned} s(d_2) &= m(1+\varepsilon)s(d_1) \\ s(d_2)p_{ml}(w|d_2) &= m \cdot s(d_1)p_{ml}(w|d_1) \end{aligned} \tag{26}$$

Below, we summarize the derivation for each case of VN-DP and VN-Okapi.

i) VN-DP. We first simplify Eq. (24) by replacing $s(d_2)$ with the terms m, ε, and $s(d_1)$ using Eq. (26), as follows:

$$\mu \left(p_{ml}(w|d_1) - p(w|C) \right) \geq m \left(\mu \left(p_{ml}(w|d_1) - p(w|C) \right) - \mu\varepsilon p(w|C) - \varepsilon s(d_1) p_{ml}(w|d_1) \right) \tag{27}$$

First, when $p_{ml}(w|d) = p(w|C)$, it is easily shown that Eq. (27) holds. Thus, we do not consider the equality case of A_1 to simplify Eq. (27).

Under the remaining cases of A_1 (i.e., $p_{ml}(w|d) > p(w|C)$), Eq. (27) is equivalent to

$$\frac{v(d_2)}{v(d_1)} \geq \left(1 - \frac{K}{|d_1|} \frac{p(w|C) + p_{ml}(w|d_1)s(d_1)\mu^{-1}}{p_{ml}(w|d_1) - p(w|C)} \right) \tag{28}$$

Under $p_{ml}(w|d) > p(w|C)$, the right-hand side of Eq. (28) is a decreasing function with respect to $p(w|C)$, with the upper bound when $p(w|C) = 0$. After replacing the right-hand side with this upper bound, the necessary condition for Eq. (28) is simplified to:

$$\frac{v(d_2)}{v(d_1)} \geq \left(1 - \frac{K}{|d_1|} \frac{s(d_1)}{\mu} \right) = 1 - \frac{K}{v(d_1)} \frac{1}{\mu} \tag{29}$$

which is satisfied if C_1 is true, regardless of the choice of parameter μ.

ii) VN-Okapi. As in the case of VN-DP, we replace $s(d_1)$ with the terms m and ε based on Eq. (26), simplifying Eq. (25) to

$$(1-b)idf(w) \geq m\left((1-b) - b\varepsilon\frac{s(d_1)}{avgs}\right)idf(w) \tag{30}$$

First, when $idf(w) = 0$, it is clear that Eq. (30) holds.
Second, when $idf(w) > 0$ (under A_1), Eq. (30) is further rewritten as

$$\begin{aligned}\frac{1}{m} &\geq 1 - \varepsilon\frac{b\cdot s(d_1)}{(1-b)avgs} \\ &= 1 - \frac{b\cdot K}{(1-b)v(d_1)\cdot avgs}\end{aligned} \tag{31}$$

which is equivalent to

$$\frac{b\cdot K}{(1-b)avgs} \geq v(d_1) - v(d_2) \tag{32}$$

Thus, LNC1 is satisfied if C_1 is true.

2) LNC2

We could straight forwardly derive that LNC2 is equivalent to C_2. To simplify the notation for the derivation, we introduce $\rho = p_{ml}(w|d_1) = p_{ml}(w|d_2)$; the equality holds because of the characteristic of PLS. We summarize the derivation of C_2 for each of VN-DP and VN-Okapi.

i) VN-DP. For VN-DP, Eq. (24) is simplified to

$$\frac{s(d_1)\rho + \mu p(w|C)}{s(d_1)+\mu} \geq \frac{s(d_2)\rho + \mu p(w|C)}{s(d_2)+\mu} \tag{33}$$

It is trivial to show that the necessary and sufficient condition for LNC2 is C_2, if A_1 holds:

ii) VN-Okapi. For VN-Okapi, Eq. (25) is simplified to

$$\rho s(d_1)\left(k\left(1-b+b\frac{s(d_1)}{avgs}\right)\right)idf(w) \geq \rho s(d_2)\left(k\left(1-b+b\frac{s(d_2)}{avgs}\right)\right)idf(w) \tag{34}$$

Using $idf(w) \geq 0$ from A_1, Eq. (34) is equivalent to C_2.

3) TF-LNC

For the sake of convenience, we introduce variables m' and ε' by putting $m' = v(d_2)/v(d_1)$ and $\varepsilon' = K/|d_2|$. According to the definition of PAR,

$$\begin{aligned}s(d_1) &= m'(1+\varepsilon')s(d_2) \\ s(d_1)p_{ml}(w|d_1) &= m'\left(p_{ml}(w|d_2)+\varepsilon'\right)s(d_2)\end{aligned} \tag{35}$$

Below, we summarize the derivation for each of VN-DP and VN-Okapi.

i) VN-DP. We first simplify Eq. (24) by replacing $s(d_1)$ with the terms m', ε', and $s(d_2)$, as follows:

$$m' \left(\frac{\varepsilon' \left(1 - p_{ml}(w|d_2)s(d_2)\right)}{\mu} + \left(p_{ml}(w|d_2) - p(w|C)\right) + \varepsilon' \left(1 - p(w|C)\right) \right) > \left(p_{ml}(w|d_2) - p(w|C)\right) \tag{36}$$

When $p_{ml}(w|d) = p(w|C)$, it is easily shown that Eq. (36) holds.
When $p_{ml}(w|d) > p(w|C)$, in the remaining cases of A_1, Eq. (36) is equivalent to

$$\frac{v(d_1)}{v(d_2)} < 1 + \frac{K}{|d_2|} \frac{\left(1 - p(w|C)\right) + \left(1 - p_{ml}(w|d_2)\right) s(d_2)\mu^{-1}}{p_{ml}(w|d_2) - p(w|C)} \tag{37}$$

For $p_{ml}(w|d) > p(w|C)$, the right-hand side of Eq. (37) is an increasing function with respect to $p(w|C)$, with the lower bound when $p(w|C)$ = 0. In addition, we can further lower the bound by eliminating $(1-p(w|d_1))s(d_2)/\mu$ because it is a positive value. Thus, we obtain the following necessary condition for Eq. (37):

$$\frac{v(d_1)}{v(d_2)} \leq 1 + \frac{K}{c(w, d_2)} \tag{38}$$

which is equivalent to C_3[16].

ii) VN-Okapi. As in the case of VN-DP, we replace $s(d_1)$ in the terms m', ε', and $s(d_2)$ using Eq. (35) simplifying Eq. (25) to

$$m' \left(\varepsilon' \left(1 - p_{ml}(w|d)\right) \frac{b \cdot s(d_2)}{avgs} + \left(p_{ml}(w|d) + \varepsilon'\right)(1 - b) \right) idf(w) \geq p_{ml}(w|d)(1 - b) idf(w) \tag{39}$$

Under A_1, because $idf(w) \geq 0$, Eq. (39) is equivalent to

$$\frac{v(d_1)}{v(d_2)} < 1 + \frac{K}{|d_2|} \frac{(1 - b) + \left(1 - p_{ml}(w|d_2)\right) b \frac{s(d_2)}{avgs}}{p_{ml}(w|d_2)(1 - b)} \tag{40}$$

A lower bound for the right-hand side Eq. (40) is obtained by eliminating (1-$p_{ml}(w|d))bs(d_2)/avgs$), which is a positive value. Therefore, the necessary condition for Eq. (40) becomes

$$\begin{aligned} \frac{v(d_1)}{v(d_2)} &\leq 1 + \frac{K}{|d_2|} \frac{1}{p_{ml}(w|d_2)} \\ &= 1 + \frac{K}{c(w, d_2)} \end{aligned} \tag{41}$$

which is equivalent to C_3 [17].

Appendix D: Analysis Result of TF-LNC under UniqLength and LengthPower

TF-LNC is true when using UniqLength and LengthPower. To prove this, let Δ_s be $s(d_1) - s(d_2)$. First, when $\Delta_s \geq K/v(d_2)$, it is equivalent to $v(d_1) \leq v(d_2)$, and thus, C_3 is satisfied. Otherwise (i.e., $\Delta_s < K/v(d_2)$), C_3 is equivalent to

$$c(w, d_2) \leq K \frac{v(d_2)}{v(d_1) - v(d_2)} = K \frac{|d_2| + \Delta_s \cdot v(d_2)}{K - \Delta_s \cdot v(d_2)} \tag{42}$$

[16]Note that Eq. (38) can contain the equality condition, because $p(w|C) > 0$ (even when $p(w|C) \to 0$)
[17] Note that Eq. (41) can allow the equality condition.

The term on the right-hand side of Eq. (42) is a decreasing function with respect to Δ_s. For the cases of UniqLength and LengthPower, $\Delta_s \geq 0$, regardless of K; therefore, Eq. (42) is satisfied if $c(w, d_2) \leq |d_2|$ (i.e., obtained by using $\Delta_s = 0$ in the right-hand side of Eq. (42)), which is true for all cases:

Appendix E: Analysis Result of TF-LNC under EntropyPower

In this appendix, for VN models using EntropyPower, we show that TF-LNC is satisfied if C_4 is true. To prove this, let Δ_s be $s(d_1) - s(d_2)$. When $\Delta_s \geq K/v(d_2)$, it is equivalent to $v(d_1) \leq v(d_2)$, and thus, C_3 is satisfied. Otherwise, C_3 is equivalent to:.

$$\Delta_s \geq \frac{K\left(c(w,d_2) - |d_2|\right)}{\left(c(w,d_2)+K\right)v(d_2)} \tag{43}$$

Eq. (43) is further simplified to:

$$\frac{s(d_1)}{s(d_2)} \geq \frac{c(w,d_2)}{c(w,d_2)+1}\frac{|d_2|+K}{|d_2|} \tag{44}$$

Using the definition of EntropyPower, we can rewrite $\log(s(d_1))$ and $\log(s(d_2))$ as:

$$log\left(s(d_2)\right) = -\sum_{w' \in d_2} \frac{c(w', d_2)}{|d_2|} \log c(w', d_2) + \log |d_2| \tag{45}$$

$$\begin{aligned} \log\left(s(d_1)\right) = \quad & -\frac{c(w,d_2)+K}{|d_2|+K}\log\left(c(w,d_2)+K\right) \\ & -\sum_{w' \in d_2, w' \neq w} \frac{c(w', d_2)}{|d_2|} \log c(w', d_2) + \log |d_2| \end{aligned} \tag{46}$$

Substituting Eqs. (46) and (45) in Eq. (44), we now obtain the following condition for TF-LNC:

$$\begin{aligned} \log\left(s(d_1)\right) - \log\left(s(d_2)\right) \;=\; & -\frac{c(w,d_2)+K}{|d_2|+K}\log\left(c(w,d_2)+K\right) \\ & +\frac{|d_2|+K-1}{|d_2|\left(|d_2|+K\right)} \cdot c(w,d_2)\log\left(c(w,d_2)\right) \\ & +\frac{1}{|d_2|+K}\left(\log|d_2| - \log s(d_2)\right) \\ & +\log\left(|d_2|+K\right) - \log\left(|d_2|\right) \end{aligned} \tag{47}$$

From the definition of TF-LNC, it is clear that once Eq. (47) holds for K = 1, then Eq. (47) also holds for every K. Therefore, here, we only consider $K = 1$. When $K = 1$, Eq. (47) is further simplified to:

$$\begin{aligned} \frac{|d_2| - c(w,d_2)}{|d_2|+1}\log\left(c(w,d_2)+K\right) - \quad & \frac{|d_2|+1-c(w,d_2)}{|d_2|+1}\log c(w,d_2) \\ & \geq \frac{1}{|d_2|+1}\log\frac{s(d_2)}{|d_2|} \end{aligned} \tag{48}$$

which leads to:

$$\left(|d_2| - c(w,d_2)\right)\log\left(1+\frac{1}{c(w,d_2)}\right) - \log c(w,d_2) \geq \log\frac{s(d_2)}{|d_2|} \tag{49}$$

Applying $(1+x)^n \geq (1+nx)$ to the first term of the left-hand side of Eq. (49), we obtain the following sufficient condition for Eq (49):

$$\log\left(\frac{|d_2|}{c(w,d_2)}\right) - \log\left(c(w,d_2)\right) \geq \log\frac{s(d_2)}{|d_2|} \tag{50}$$

which is rewritten as

$$2\left(\log|d_2| - \log c(w,d_2)\right) \geq \log s(d_2) \tag{51}$$

Therefore, we finally obtain the condition:

$$s(d_2) \leq \left(\frac{|d_2|}{c(w,d_2)}\right)^2 \tag{52}$$

which is equivalent to C_4.